ABSTRACT

Title of Document:	FREUD, MODULARITY, AND THE PRINCIPLE OF CHARITY
	Joel Evans Gibson, Ph.D., 2010
Directed By:	Professor Georges Rey, Department of Philosophy

Within the philosophy of mind, a ‘hermeneutical’ tradition sees psychology as discontinuous with natural-scientific domains. A characteristic ingredient of this tendency is ‘normativism’, which makes obedience to rational norms an *a priori* condition on agency. In this thesis, I advance an argument against normativism which trades on the notion of a psychological module. Specifically, I show how modules can be envisioned which, because of their high degree of irrationality, challenge the normativist’s principle of charity. As an illustration, I describe such a module that incorporates key features of the Freudian ‘id’, and I suggest that Freudian theory generally puts pressure on charity constraints. In sum, I seek to substantially undermine the hermeneutical view of the mind by attacking one of its central pillars. In Chapter 1, after setting out the essential features of hermeneuticism, I sketch the historical background of recent normativism by considering Quine’s employment of charity in his theory of meaning and mind. Most centrally, I reject pragmatic and heuristic readings of Quinean charity in favor of one that sees it as a constitutive

constraint on attribution. In Chapter 2, I begin to clarify the content of Davidsonian charity, against which—in the first instance—my argument levels. I identify Maximization and Threshold Principles in Davidson's early papers, contrast Davidsonian charity with Richard Grandy's Principle of Humanity, and rebut typical arguments for charity principles. In Chapter 3, after identifying two additional Davidsonian charity principles (a Competence and a Compartment Principle) and describing the conception of a module figuring in my argument, I present my argument in schematic form. Then I critique attempts to rebut my argument through excluding modular processes from the scope of normativism (notably, via a personal-subpersonal distinction). In Chapter 4, I develop my argument in detail by describing a module that embodies basic forms of Freudian wish-fulfilment and demonstrating how it violates charity principles. Further, I rebut possible objections to my use of Freudian theory. In Chapter 5, I canvass various models of Freudian phenomena more generally and suggest that a version of my argument can be run with respect to such phenomena too.

FREUD, MODULARITY, AND THE PRINCIPLE OF CHARITY

By

Joel Evans Gibson

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Ph.D.
2010

Advisory Committee:
Professor Georges Rey, Chair
Professor Peter Carruthers
Professor Paul Pietroski
Associate Professor Michael Morreau
Associate Professor Michael Dougherty

Dedication

To my mother and in memory of my father.

Acknowledgements

Needless to say, I am endebted to Georges Rey and the other members of my thesis committee for their help and encouragement in all phases of this—truth be told—long and arduous project. But, more than that, I am grateful to them for opening up to me at least a *glimpse* into the rich world of ideas that they so fully inhabit.

Table of Contents

Chapter One: The Historical Background of Normativism

Introduction

That minds—at least human minds—are largely rational has long been a methodological presupposition of the social sciences, seeming to offer a foundation on which the social sciences might be grounded as genuinely explanatory and predictive disciplines. But a question can be raised as to the status of this presupposition. Within the philosophy of mind, an influential tradition—numbering among its adherents such philosophers as W. V. O. Quine, Donald Davidson, Daniel Dennett, Marcia Cavell, and Jennifer Hornsby—has viewed the rationality of (human) minds not merely as an empirically warranted generalization. Rather, philosophers within this tradition have championed the much stronger view, called 'normativism', according to which a large measure of obedience to the norms of rationality is a logically necessary, *a priori* condition for the possession of propositional attitudes by an agent.[1] It is this position against which I argue in this thesis.

More specifically, I shall advance an argument which trades on a notion which has enjoyed considerable and fruitful currency within psychological theory in recent years, namely, that of a 'module'. A module, as contemporary psychological theory understands the notion, is—roughly—any isolable functional component of the mind. Although this terminology is fairly new, the concept itself is not. In fact, I wish to suggest, Sigmund Freud's conception of the part of the personality which he labels the 'id' is, in all essentials, that of such a module. Indeed, it is the conception

[1] The correlative injunction to interpret agents as obeying such norms, in turn, is widely referred to as the 'principle of charity'.

of a module which does not hew to the rational norms adherence to which normativism regards as a condition for the possession of a mind. But given such a clash with normativism, the scientific possibility of Freud's conception of the id, I suggest, gives reason to call into question normativism's insistence on such a condition.[2] Thus, the example of Freud's id illustrates a tendency of modular theory to subvert normativism. In fact, the sort of argument I make by drawing on Freudian theory could plausibly be made by bringing to bear other bodies of psychological theory.[3]

In broad outline, my plan is the following: I first treat the origins of normativism in Quinean philosophy then present a detailed picture of more recent normativism, especially, the—in some measure—canonical form of it championed by Davidson, which, in large measure, represents the target of my attack. After clarifying the conception of modules that figures in my argument, I describe a module embodying processes of Freudian infantile wish-fulfilment and indicate how its irrationality presents a challenge to normativism. Last, I consider whether an analogue of my argument can be made relying on the wider range of phenomena described by Freudian theory.

In this first chapter, more particularly, I shall sketch the philosophical-historical background of recent normativism by considering Quine's employment of a principle of charity. For although Quine credits Neil Wilson with coining the term, Quine's use of the *concept* actually antedates Wilson's. Certainly, it is in the context

[2] It should be emphasized that my argument does not rely on the actual *truth* of Freudian theory. Rather, it depends merely on the bare scientific possibility of important elements of it.

[3] See the conclusion of this thesis for one suggestion as to an alternate version of the argument using more contemporary psychological theory.

of Quinean philosophy of language and mind that charity begins to assume its characteristic shape. Moreover, it is Quine’s use of the concept that directly (and decisively) influences the views of latter-day normativists such as Davidson and Dennett.

However, since there exists no consensus among commentators about the proper interpretation of Quinean charity, I shall canvass the various interpretations on offer and—somewhat tentatively—defend that which strikes me as most plausible. Despite the tentativeness of my conclusions in this regard, the process of sifting through the various interpretations of Quinean charity will allow me to clarify the role(s) which charity plays within Quine’s philosophy and, by extension, those of his successors. Most importantly, the resulting interpretation of Quinean charity will, in later chapters, serve as a foil against which the distinctive features of more recent normativism may be thrown into sharper relief. For despite its kinship with Quinean normativism, it will be seen to differ from it in important respects as well.

But to gain a proper appreciation of normativism, both Quinean and recent, it will be instructive first briefly to consider the relation of normativism to a broader tendency within the philosophy of mind which, in a sense I shall presently explain, may be characterized as *hermeneutical*. For normativism goes hand-in-hand with a certain general standpoint about the status of the mental. Moreover, seeing normativism against the backdrop of this broader philosophical stance clarifies the significance of the argument against normativism which I shall develop in later chapters. To a great extent, my project in this thesis should be understood as an assault on this hermeneutical view of the mind via an assault on one of its characteristic expressions, namely, normativism.

Hermeneuticism

In philosophic connections, the word ‘hermeneutical’ (and its cognates) has its home in the first instance within the Continental tradition, where it is closely associated with the philosophies of figures such as Heidegger and Gadamer. But it is Wilhelm Dilthey who first formulates the central hermeneutical doctrines. Foremost among them is the view that *Natur*- and *Geisteswissenschaften* (that is, natural and human sciences) demand distinctive methods, respectively. The method proper to the human sciences Dilthey dubs ‘*verstehen*’ (‘understanding’), which contrasts with the method of the natural sciences, which he designates ‘*erklaren*’ (‘explanation’).

Significantly, this hermeneutical dichotomizing of natural- and human-scientific methods, though particularly prevalent within Continental philosophy, has counterparts within Anglo-American philosophy, and specifically philosophy of mind. As Rey (2001) makes clear, there is no shortage of prominent analytic philosophers who have held that intentional states do not submit to natural-scientific methods. Among those whom Rey cites are Wittgenstein, Gilbert Ryle, Thomas Nagel, Jennifer Hornsby, John McDowell, Colin McGinn, Jaegwon Kim, David Lewis, Donald Davidson, and Daniel Dennett. It is this general view of the mind, then, regardless of whether its proponents are Continental or analytic, that I designate as ‘hermeneutical’ in the following; and, notably, as will become clear subsequently, Quine’s view of the mind is hermeneutical in just this sense.

To get a proper appreciation of this view of the mind, it will be helpful to examine Rey's discussion in the article cited of what he terms the 'insularity of folk psychology'. Rey quotes there the following passage of Jennifer Hornsby's, wherein she advocates a conception of folk psychology of this sort:

> If 'folk psychology' is construed by analogy with 'folk physics' or 'folk linguistics', then it carries the implication that folk psychology is the perhaps defective version of a subject matter that others (physicists, linguisticians [sic]) study with more appropriate methods than the folk. The implication is to be shunned: *we ought not to assume at the outset that the basis of our everyday understanding of one another is susceptible of correction and refinement by experts in some specialist field where empirical considerations of some non-commonsensical kind can be brought to bear*. (Hornsby 1997c, 3-4, as quoted in Rey 2001, 104)

The view of folk psychology as insular, then, amounts to the claim that scientific methods have nothing even potentially to contribute either to the emendation or amendment of our folk psychology. It is an autonomous domain unto itself, hewing to its own methods and answerable only to its own (narrow) evidential bases.[4]

It will be helpful briefly to consider how this conception of the insularity of folk-psychology relates to what I am calling the hermeneutical conception of psychology. First, the two conceptions *seem* to differ in that, whereas insularity posits a contrast between science and non-science (folk psychology), hermeneuticism distinguishes, rather, between two kinds of science, natural and human. Granted, one can perhaps conceive of drawing a threefold distinction between non-scientific,

[4] Cf. (Cherniak 1986, 3-5) where Cherniak uses the phrase "the autonomy of the mental" to describe the view of folk psychology as insular. The metaphor may differ but the conception expressed is the same. This sense of autonomy, however, should be contrasted with a sense that arises in discussions of how psychology relates to more basic fields such as neurobiology and physics. The view that psychology is autonomous vis-à-vis such fields is the view that it is not *reducible* to them. The "autonomy of the mental" in the sense of insularity, by contrast, although entailing non-reducibility, is a broader doctrine, denying the relevance of natural-science methods to the mental (or at least to folk psychology) altogether. Many adherents of non-reducibility, Rey included, emphatically reject autonomy as insularity.

natural-scientific, and human-scientific approaches to the mind. But, in fact, as will emerge subsequently, hermeneuticists tend to see the *geisteswissenschaftliche* method as largely continuous with folk method. The human sciences are at most seen as somehow refining or *extending* commonsense approaches to their subject matter rather than as truly distinctive. So the bruited contrast between insularity and hermeneuticism is more apparent than real.

A second apparent difference between insularity and hermeneuticism is that, whereas the former merely concerns folk-psychological states, the latter is ostensibly a doctrine about all psychological states. In fact, as Rey points out, some insularists appear to countenance the possibility of a cognitive science, proceeding according to natural-scientific methods, so long as this is understood to address a distinctive domain of psychological states isolated taxonomically and causally from the domain of familiar folk-psychological states. Accordingly, it may make sense to distinguish a *wide* hermeneutical view, which excludes all intentional states from the purview of natural-scientific methods, from a *narrow* hermeneuticism, which does so only for folk-psychological states, implying no commitment with respect to other psychological states, if any. Strictly, then, insularity amounts to the latter view. In fact, however, aside from some more recent adherents of hermeneutical views, adherents of hermeneutical views generally seem either not to have considered the possibility of a separate cognitive psychology, or to have disallowed its possibility—a fact which tends to diminish the import of the distinction between wide and narrow varieties of hermeneuticism. So the notion of insularity corresponds quite closely to that of the hermeneutical.

But more needs to be said by way of clarifying the hermeneuticist's main contention: What does it mean to say that intentional states are not amenable to study by natural-scientific methods, and what might lead one to adopt such a view? In order to bring the essence of hermeneuticism into focus, it will be instructive to consider the main features of Dilthey's (and Max Weber's) early hermeneuticism.

Early Hermeneuticism

Dilthey and Weber posit a mode of understanding human activities which differs from that in which non-human, natural phenomena are to be explained. First, one understands an individual or groups' behavior in terms of their subjective states (or "meanings" in hermeneutical jargon) (Nagel 1979, 480-81) rather than in terms of physical or biological states lacking content. Moreover, the human-sciences' distinctive subject-matter dictates, in Dilthey and Weber's view, a distinctive method. On the standard deductive-nomological model (cf. Hempel 1966), the natural sciences paradigmatically explain a phenomenon by citing one or more laws which, together with statements of attendant circumstances, deductively entail a statement describing the phenomenon's occurrence. Thus, explanation in the natural sciences requires the antecedent formulation and confirmation of reasonably strict general laws which subsume the phenomenon to be explained.

In understanding human activity, however, in Dilthey's view, one is able to short-circuit the laborious appeal to laws which characterizes explanation in the natural sciences. For Dilthey posits a distinctive cognitive faculty of "empathy" which affords one an intuitive understanding of others' activity. This faculty allows one to come to an understanding of another's behavior "from the inside" through a process of imaginatively “reliving” (*nacherleben*) their mental states. In this way, I come to a grasp of the springs of their behavior through imaginative access to what would lead *me* to behave so if I were “in their shoes,” that is, similarly situated.[5]

Further, on Dilthey’s classical hermeneuticist conception, the physical and human sciences differ in the additional respect that whereas the former seek to explain phenomena in terms of their causes, the latter cite subjective states which are *non-causal* as the grounds for human beings’ behavior. On this view, the knowledge of “the inner life of another person” is “not a knowledge of causal connections but rather of a network of meanings, analogous to the network of meanings by which I understand myself” (Phillips 1996, 62). Dilthey’s hermeneuticism, then, contains several features which are characteristic of hermeneutical views of the mind: It understands human beings’ activity by ascribing intentional states to them; it employs a non-natural-scientific method (“reliving” through a faculty of “empathy”) by which such understanding is acquired; it yields explanations of behavior which lack the covering-law form typical of the natural sciences; and it regards the intentional states through which it interprets human activity as non-causal grounds of the latter. But not every feature of Dilthey’s hermeneuticism is essential to that philosophical tendency as I understand it. To see this, it will be useful to consider some of these features in a bit more detail.

Features of Hermeneuticism

First, as I have said, hermeneuticism is at bottom the view that the human- and natural-sciences have distinctive methods. But some delicacy is required with respect

[5] The modern view that our understanding of others proceeds by way of simulation rather than on the basis of theory is of a piece with Dilthey’s view. On simulation see (Carruthers and Smith, 1996).

to the word ‘method’. There is a sense in which it is trivial that the human and natural sciences have different methods. For example, the peculiar subject-matter of the former (attitudes, beliefs, etc.) permits the use of survey-techniques which have no application in the natural sciences. In fact, a more careful formulation of the definition of hermeneuticism puts it in terms, not of ‘method’ but of ‘methodology’: hermeneuticism is the view that the human- and natural-sciences have distinctive *methodologies*. That is, on the hermeneutical view, there are distinctive standards by which human-scientific hypotheses are to be confirmed or disconfirmed. As Nagel puts it, “The crucial question . . . is whether” ascriptions of subjective states in the human sciences “involve the use of logical canons which are different from those employed in connection with the imputation of ‘objective’ traits to things in other areas of inquiry” (1979, 481). This is what I take to be essential to hermeneuticism.

Accordingly, it is not the employment of ‘reliving’ or ‘empathy’ *per se* which makes Dilthey’s view of the human sciences hermeneuticist. Indeed, such processes may very well have a heuristic role to play in the discovery of human-scientific explanatory hypotheses (Nagel 1979, 484), without thereby possessing a properly methodological significance. What makes Dilthey’s view hermeneuticist is that he apparently *does* assign ‘reliving’ and ‘empathy’ this significance: the origin of an explanatory hypothesis through this cognitive channel is confirmatory of it. That is not to say, however, that all hermeneuticist views locate the methodological peculiarity of the human sciences precisely where Dilthey does, in a faculty of “reliving” or “empathy.” There are numerous proposals for what this peculiarity consists in.[6] But what unites hermeneutical views is the epistemological point that they take the human sciences to be methodologically distinctive in one or another respect.

This means that other features often associated with hermeneuticism—such as Dilthey's metaphysical view of intentional states as non-causal—are merely secondary, contingent ones. The view that mental states are non-causal has traditionally been widely held among proponents of hermeneuticism. This view typically takes the form of maintaining that mental states are *reasons* rather than causes, that the relations that exist among mental states are logical rather than causal. But under the pressure of Donald Davidson's forceful argumentation (Davidson 1980a), most have abandoned the view that reasons cannot be causes. So latter-day hermeneuticists are less inclined to share Dilthey's view of intentional states as non-causal. Indeed, as will emerge later, Davidson's view of the mental is itself hermeneuticist despite his holding that mental states are causal. So it is not essential to hermeneuticism to see mental states as non-causal.[7]

As I have indicated, Dilthey holds that the human sciences dispense with strict laws and explanations based on them. The hermeneuticist's stress on intentional states as reasons lends a certain initial plausibility to this conception. For as Davidson points out, reason-explanations of the sort with which our everyday folk-psychological explanatory practices are replete do in fact lack the covering-law form typical of the natural sciences (1980a). In fact, if the human sciences relied only on

[6] For a valuable survey of several different such proposals see (Erwin 1996, 8-41).

[7] I intend this as a *prima facie* claim. It may be that, in ways that are not immediately apparent, a well-considered hermeneuticism would commit one to the view that mental states are non-causal. Indeed, several commentators on Davidson's philosophy of mind argue that his hermeneutical outlook ill comports with his view of mental states as causal (Antony 1989, Child 1994, Evnine 1991).

such explanations and eschewed covering-law explanations altogether, then that would seem sufficient to guarantee the irrelevance of natural-scientific methodology to them. For natural-science hypotheses are either themselves laws or are tested by drawing out their entailments in conjunction with laws and observing whether they obtain. Strict laws and the predictions and explanations based on them appear to be part and parcel of natural-scientific method. So it does appear necessary for the hermeneuticist to deny their relevance to the human sciences if he is to maintain their methodological autonomy.

Aside from the above features of Dilthey's hermeneuticism, there is some temptation to see a close connection between hermeneuticism and an anti-realism or instrumentalism with respect to the mind. Certainly, some prominent Anglo-American adherents of hermeneutical views—notably Daniel Dennett (see Rey 1994), Colin McGinn[8], and, as we will see, Quine—have assigned the mind a secondary ontological status. But there is no indication that Dilthey was an anti-realist; and Davidson explicitly characterizes his theory as realist. So pending compelling argument, one should avoid concluding that hermeneuticists are necessarily committed to something less than a realism with respect to the mind.[9]

Hermeneuticism's Appeal

Early figures like Dilthey and Weber aside, one might wonder why so many prominent contemporary philosophers of mind have been attracted to hermeneuticism. Rey (2001) discusses some considerations which, he suggests, have

[8] Rey (2001, 106) cites Colin McGinn (1991), where McGinn claims "our mental concepts are happily superficial."

[9] It should also be pointed out that in denying the applicability of natural-scientific methods to the study of the mind, the hermeneuticist is not thereby committing himself to a denial of physicalism, the view that all that exists belongs to the physical world.

tempted philosophers within the Analytic tradition to the view that folk psychology is insular, that is, in effect (cf. p. 5 above), the hermeneutical conception of psychology. First, there is a functionalism of the sort advocated by David Lewis (1972). On this view, sometimes referred to as "folk functionalism" (cf. Rey 1997, 185), mental vocabulary is defined by Ramsifying over the set of "platitudes" concerning mental states "which are common knowledge among us" (Lewis 1972, 212). Now, as Rey notes (2001, 105), such an approach has the effect of insulating folk psychology from scientific inquiry at least to this extent: it views mental concepts as *terra cognita*, whose essences are patent, and not as natural kinds whose essences, as on a Kripkean scientific-essentialist conception, are to be illuminated by empirical investigation. So natural-scientific methods have no role to play for Lewis in defining our mental concepts.

But it should be noted that such folk functionalism falls short of altogether entailing insularity. For suppose that mental concepts are defined along the lines Lewis suggests. That would mean that mental states would be picked out by our commonsensical knowledge of such states, but it need not be the case that that knowledge was exhaustive: there could be additional facts concerning those mental states discoverable by natural-scientific methods. So even if those methods would not on Lewis's account play a role in defining mental concepts, they could still in principle serve to expand our knowledge concerning those mental states. Hence, folk functionalism does not entail the complete insularity of folk psychology.

A second source for the insularity view discussed by Rey (2001, 107) is the view that mental concepts have their home in the context of reason-explanations.

These are “explanations in which things are made intelligible by being revealed to be, or to approximate being, as they rationally ought to be” in contrast to covering-law explanations where things are shown to be instances of “how things generally tend to happen” (McDowell 1985, 389, quoted in Rey 2001, 107). Since the rational norms governing these states are presumably *a priori*, no role remains for an empirical psychology.

This view, if correct, does have the implication that folk psychology would be insulated from scientific investigation, for the latter, as I have suggested (see p. 11 above), stands or falls with the existence of laws and explanations couched in terms of them. But though an important role of mental concepts is undoubtedly the one highlighted by this line of argument—their role in rationalizing explanations—one justifiably seeks reason why one should accept that mental states play *only* that role. Why should one accept that there are no laws involving mental states and, therefore, no explanations in terms of them?

A third route to insularity discussed by Rey—one very pertinent to my project—tackles this question head-on. This route is Davidson’s appeal to normativism in arguing the “anomalousness of the mental,” the doctrine that there are no strict psychophysical laws. The locus classicus for Davidson’s argument is his “Mental Events” (1970). The proper interpretation and assessment of the argument are matters concerning which there has been considerable debate. But if the argument carries, then it *would* indeed seem to have the consequence of undermining the scientific pretensions of psychology. For as Jaegwon Kim points out, “Science is supposed to be nomothetic . . . so that where there can be no laws there can be no science and . . . we have no business pretending to be doing science" (Kim 1993e, 194-96, quoted in Rey 2001, 111). So, in this way, normativism may entail a hermeneutical conception of the mind and, therefore, the project of refuting normativism acquires significance as part of a defense of the possibility of a scientific approach to the mind.

Quine as Normativist

With this sketch of the broader philosophic context of normativism in place, I now turn to Quine as the immediate precursor of much contemporary normativist thought. I begin with a discussion of Quine's general conception of language and the mind before turning specifically to the role of the principle of charity in his thought. The relevance of the hermeneutical conception of the mind in his philosophy will be apparent at several points.

Quine's Interpretationism

Quine's general approach to language and to mind, I think, can be characterized as *interpretationist*. Although it is Davidson and Dennett whose views are more commonly described as such, the designation seems to fit Quine as well, who in this—as in other respects—is their forerunner. With respect to the mind, "Interpretationists regard an agent's being endowed with a mind as a matter not of that agent's possessing a particular material make-up . . . or a particular kind of internal organization" (Heil 2004, 10), but as a result of the agent's being so interpretable on the basis of their behavior. "The interpretationist thought is that we can give an account of the circumstances under which it is true that *S* believes that *p* by considering the circumstances under which *S* could be interpreted as believing that *p*, on the basis of what she says and does" (Child 1994, 3). But a view which grounds an account of the meaning of an individual or group's language in how they can be interpreted on the basis of their behavior will count as interpretationist as well.[10] And Quine indisputably provides an account of language of this general sort.[11]

Quine's account of linguistic meaning, which he develops in *Word and Object* (1960), is based on the concept of radical translation.[12] Quine gives an account of meaning by describing the constraints which, he holds, govern the translation of the sentences of "the language of a hitherto untouched people" into the sentences of an interpreter's language.[13] Such translation is "radical" because it proceeds without prior knowledge of the language or assistance from third parties in possession of this knowledge. The translator is forced to rely on whatever evidence is provided by the behavior of his or her "informants." Specifically, it is the "stimulus conditions" of utterances, in effect, the circumstances in which they are elicited that are to guide the translator in their translation.

If the interpreter can discover the expressions for assent and dissent in the informants' language, he will find that some sentences are such that informants' assent and dissent to them follows a uniform pattern: each informant assents (or dissents) to a given such sentence in precisely the same circumstances, where

[10] In fact, as William Child notes (1994, 13), Davidson—the arch-interpretationist—actually reserves the designation 'interpretation theory' to the ascription of meaning to language. He uses 'decision theory' to denote the project of making sense of an individual's propositional attitudes.

[11] I consider below to what extent Quine can be reckoned an interpretationist with respect to the mind as well.

[12] My summary of Quine's views on language is heavily indebted to Hookway (1998).

[13] Quine describes the interpreter as compiling a "manual" which will allow translation of the natives' utterances into his own language. However, Quine is vague about what such a manual would look like in detail. He does write (1960, 68) that the linguist in compiling a manual will segment natives' utterances into words and phrases, which he will match with words and phrases of his own language to serve as "analytical hypotheses." So a manual, it appears, will rather resemble a dictionary. As will appear later, Donald Davidson has a much more developed conception of what such a manual would require.

sameness of circumstances is a matter of the sameness of the sensory stimulation to which they are exposed. These sentences Quine refers to as “observation sentences,” and the ordered pair consisting of the set of stimuli that elicit an affirmative response to such a sentence and the set of stimuli that evoke a negative response to it Quine calls the observation sentence’s “stimulus meaning.”

It is these observation sentences which offer the linguist entrée into the natives’ language and represent a key constraint on translation for Quine. For an acceptable manual of translation, Quine holds, must correlate with each observation sentence of the natives’ language an observation sentence of the linguist’s language, and, in fact, one which is “stimulus synonymous” with it, in the sense that it possesses the same stimulus meaning.

Quine’s classic illustration involves the word (or sentence) ‘Gavagai’, which an English-speaking linguist observes natives to utter frequently in the presence of rabbits. Moreover, when he asks natives “Gavagai?,” he observes that they assent (and dissent) in precisely those stimulus conditions in which English speakers do to the sentence “‘Rabbit’ (or ‘Lo, a rabbit’)” (1960, 29). That is to say, the native’s “Gavagai” and linguist’s ‘Rabbit’ are stimulus synonymous.

Two points should be noted about stimulus meaning and stimulus synonymy, respectively. First, the notion of stimulus meaning cannot be taken as providing some sort of gloss on the ordinary notion of meaning. For only a small subset of sentences will possess a stimulus meaning at all, inasmuch as most sentences will not record observations of one’s immediate environment, and so one’s assent or dissent to them will depend partly on such things as what background beliefs one has. Second, it is not enough to ensure that ‘Gavagai’ is correctly translated as ‘Rabbit’ that the two sentences be stimulus synonymous. Rather, the correctness of such a translation is vouchsafed only in virtue of its being dictated by a complete translation manual that yields that translation—and there are other constraints on the correctness of such a manual in addition to preservation of stimulus meaning across languages.

Quine notes that it should be possible to identify those expressions of the natives’ language that function as logical connectives. So, e.g., a negation-operator can be identified as any expression which “turns any short sentence to which one will assent into a sentence from which one will dissent, and vice versa” (1960, 57).[14] Quine requires that a translation manual reflect this in that native expressions which correspond in this way to truth-functional operators are to be so rendered. Quine notes, further, that it is possible to identify “intrasubjective stimulus synonymy of sentences” (1960, 68) by a native’s assenting and dissenting to each in precisely the same circumstances. Quine stipulates that the analytic hypotheses constituting a translation manual should yield translations of sentences which are stimulus-synonymous for natives with English sentences that are stimulus-synonymous for English speakers. Again, some sentences, which Quine labels “stimulus-analytic (-contradictory),” are assented (dissented) to by natives regardless of stimulus. Quine requires that a translation manual yield translations of them that are stimulus-analytic (-contradictory) for English speakers.[15]

Such are the lineaments of Quine’s interpretationist theory of linguistic meaning. It is a theory from which Quine draws far-reaching implications. For

[14] Quine’s view of the role of logical connectives is bound up with his views about charity, which I discuss below.

[15] This requirement too has implications for Quine’s conception of charity, which I discuss below.

Quine maintains that even when all possible evidence is taken into account, there will be multiple manuals that fulfill the above four constraints which differ substantively among themselves in that they map particular native sentences onto English sentences which are not stimulus-synonymous for English speakers. So, e.g., the term 'gavagai', which in virtue of its stimulus-meaning when used as a sentence, seemed appropriately translated as 'rabbit', might—so far as stimulus-meaning is concerned—just as well be rendered 'rabbit stage' or 'undetached rabbit part', since the stimulus-meaning of these expressions is the same. But 'Lo, there is a rabbit stage' is not stimulus-synonymous with 'Lo, there is a rabbit'. Quine asserts that these various translations "could doubtless be accommodated by compensatory variations in analytical hypotheses concerning other locutions, so as to conform equally to . . . all speech dispositions of all speakers concerned" (1960, 72). Quine's point is that though one might hope to discriminate between, say, 'rabbit' and 'undetached rabbit part' as a rendering of 'gavagai' by natives' readiness to assent in the presence of a rabbit when the word appears in a context naturally translated 'There is only one X present', one can preserve the rendering as 'undetached rabbit part' by preferring an alternative rendering of that context. The upshot is that the sources of evidence for translation expressed in the four constraints "woefully under-determine the analytical hypotheses" (1960, 72). Thus, Quine holds, there are multiple adequate but incompatible translations, none of which can lay claim to greater correctness than its fellows. This is Quine's famous doctrine of the indeterminacy of translation.[16]

[16] For critical discussion of whether such indeterminacy really follows from Quine's account of translation see (Hookway 1998, Ch. 9).

Underpinnings of Quine’s Theory of Meaning

Before I address the implications Quine draws from that doctrine, I want to examine its theoretical underpinnings. For those implications will only be as strong as the foundations of the theory of meaning which entails them. Quine’s views about language are usefully seen against the backdrop of his epistemological and ontological orientation. Quine is an empiricist, although he represents an empiricism which differs in important ways from that of his logical positivist predecessors. In particular, in (1980) he both challenges the orthodox distinction of truths into analytic and synthetic as well as the verificationist theory of meaning, dubbed ‘translationism’ by Quine, which holds that the meaning of a sentence is the set of observations which confirm it. This conception of meaning founders on the Duhemian confirmation-holism to which Quine ascribes. Moreover, empiricist that he is, Quine will have no truck with the Fregean view of meanings as abstract entities. A tenable theory of meaning for Quine needs to comport with the ontological austerity that empiricism imposes.

But it is the sciences, and especially physics, which for Quine are the exemplars of empiricist epistemological canons. Thus, Quine is led on the basis of his empiricism to a physicalist ontology. As Hookway points out (1998, 71-72), Quine subscribe to a ‘determinationist’ doctrine according to which physics is the fundamental science, inasmuch as all facts supervene on physical ones. Moreover, Quine views this as warrant for concluding that only physical objects genuinely exist, that only physics represents a genuinely factual discourse. Since, he thinks, non-basic sciences such as geology and traditional mentalist psychology (and possibly even biology and chemistry) are not reducible to physics, they are not strictly factual. Only claims that can be cashed in rigorously physical terms can count as factual.[17]

Not surprisingly, then, Quine eschews mentalist psychology, preferring a behavioristic and physiological psychology which more closely toes the line of his austere physicalism.[18] And this behaviorism is much in evidence in Quine's theory of radical translation. For the only evidence to which the interpreter can appeal in constructing a translation manual is behavioral. So translatability is constituted in behavioral terms. Indeed, Quine's central notion of stimulus meaning is itself intended as something of a respectable behavioristic *ersatz* for what he regards as a hopelessly vague pre-theoretic notion of meaning.[19]

[17] Hookway observes (1998, 50-54) that Quine's physicalist realism is more characteristic of his later writings. In his earlier works, by contrast, he seems by and large to espouse an anti-realist pluralism or relativism on which no scheme is fundamental or can lay claim to strict factuality.
Moreover, as Hookway notes (1998, 76-77), Quine's inference of physics' exclusive factuality on the basis of his determinationism may be too quick. For though physics may describe the fine structure of reality, there may be broader structural features of reality that only the special sciences reveal (cf. Fodor 1974).

[18] In some respects, however, the appeal of behaviorism for Quine is puzzling. After all, Quine's confirmation holism negates one of the main motivations for behaviorism. For methodological behaviorism in psychology is based on the view "that only what is publicly observable is a fit subject for science" (Heil 2004, 65). Confirmational holism loosens the connection between entities and the observational evidence which confirms their existence, thereby accommodating unobservable theoretical entities—rightly—within the purview of science. As Heil astutely observes (2004, 52), "Philosophers impressed by behaviorism in psychology sometimes failed to appreciate the extent to which the behaviorist conception of mind was the product of a contentious philosophical conception of scientific method. Ironically, the roots of that conception lay in a positivist tradition that many of these same philosophers would have found unappealing." This applies *par excellence* to Quine. Indeed, Quine is himself one of the chief demolition-experts of that very positivist tradition that undergirds psychological behaviorism.

[19] Quine's reference to "the conceptual slough of meaning" (1960, 43) vividly betrays his attitude to that pre-theoretical notion.

Implications for Meaning and Mind

Now as noted, the theory based on such stimulus meaning seems to leave translation indeterminate.[20] Quine grants that some manuals will be preferable for pragmatic reasons. For example, choosing to translate natives in such a way that the statements they make are rendered by statements *we* would make in similar circumstances may make mutual cooperation easier. Again, the relative simplicity of the translations a manual yields may recommend it.[21] But such pragmatic considerations are not substantive constraints; they do not reduce the range of factually correct translations. The indeterminacy remains unabated.

This indeterminacy Quine takes to warrant skepticism about meaning.[22] As Hookway observes, by contrast with the translationism that Quine subverts in (1980), Quine's account of meaning may appear to be a form of semantic holism, a view in which the whole language is in some sense the unit of meaning. Sentences (and words) carry what meaning they have only via the rendering they receive along with all other sentences of the language in the context of radical translation. But, in fact, Quine's view seems to be that once the positivist, translationist conception of meaning is undermined, there is not "much to say about meaning at all" (Hookway

[20] Indeed, the renderings of the same word in alternate translation manuals can even fail of co-referentiality (e.g., 'rabbit' and 'rabbit-stage' for 'gavagai'). Quine refers to this consequence of his scheme of radical translation as 'the inscrutability of reference'.

[21] Indeed, pure aesthetic reasons may lead one to favor one manual over others. (In fact, simplicity may be one relevant aesthetic value.) One can even envision emotional reasons: Choosing to render the bulk of natives' statements as about objects rather than object-stages or undetached parts of them (cf. 'gavagai') might mitigate a sense of alienation towards the natives.
More seriously, on Hookway's view, charity for Quine is merely one pragmatic consideration among others bearing on the choice of manuals. It is not a constitutive constraint on translation. I return to this reading of Quinean charity below.

[22] I.e., a constitutive skepticism about the very existence of meaning, not an epistemological skepticism concerning whether meaning is knowable (cf. Miller 1998, 132).
As will emerge later, Davidson by contrast does not take indeterminacy of his radical interpretation to support such skepticism.

1988, 166). The indeterminacy of translation spells doom for the last best hope for a respectably empiricist account of meaning.

But more than this, Quine draws equally radical implications from his doctrine for the mind. In our everyday practice, we explain human behavior through propositional attitudes such as belief and desire. Such states are identified largely through their content, which we identify through that-clauses (‘Ptolemy believed *that* the sun revolved around the earth’). But given the indeterminacy of translation, such attributions of propositional attitude seem to ascribe no *definite* content. The native’s assent to ‘Gavagai’, e.g., is equally well interpreted as an expression of his belief that there is a rabbit before him, or the belief that there is an undetached rabbit-part before him. There is no fact-of-the-matter which is the native’s true belief, Quine concludes. Thus, the indeterminacy of translation comes to infect intentional psychology as well as language.[23]

Quine observes that Franz Brentano had argued for the thesis that “there is no breaking out of the intentional vocabulary by explaining its members in other terms” (1960, 220). That is, Brentano had argued that it is impossible to give an account of intentional psychology (and other intentional domains) in purely physicalist terms, and infers on this basis that the mind is an autonomous realm. Quine himself accepts Brentano’s premise (the indeterminacy of translation establishes as much) but draws a different conclusion from it:

> One may accept the Brentano thesis either as showing the indispensability of the intentional idioms and the importance of an autonomous science of intention, or as showing the baselessness of intentional idioms and the

[23] “Evidently, then, the relativity to non-unique systems of analytical hypotheses invests not only translational synonymy but intentional notions generally” (1960, 221).

> emptiness of a science of intention. My attitude, unlike Brentano's, is the second. (1960, 221)

Given Quine's prior commitment to physicalism, which for him involves the notion that only what can actually be cashed in physical terms can lay claim to factuality,[24] he feels entitled to infer that psychology is "baseless." He emphasizes, however, that he is not urging the abandonment of intentional idioms; he acknowledges their practical utility. But he insists, "If we are limning the true and ultimate structure of reality, the canonical scheme for us is the austere scheme that knows . . . no propositional attitudes but only the physical constitution and behavior of organisms" (1960, 221). So Quine appears to infer a quite thoroughgoing anti-realism with respect to the mind on the basis of the indeterminacy of translation.

Criticisms of Quine's Argument

Naturally, Quine's doctrine of the indeterminacy of translation has not sat well with many, and, as Hookway notes (1998, 144-45), it is tempting to see Quine's argument for it rather as a *reductio ad absurdum* of one or more of its premises. In particular, Quine's behavioristic restriction of the possible evidence for translation largely to the facts about observation sentences' stimulus-meaning has seemed vulnerable. Noam Chomsky observes that it is not surprising that incompatible translations are obtainable when evidence is restricted so, but "It is less obvious that there are incompatible hypotheses such that no imaginable evidence can bear on the choice between them" (Chomsky and Katz 1974, n7). In fact, as Hookway notes, there would appear to be a wealth of other sorts of evidence bearing on the selection

[24] Essentially, i.e., he regards physicalism as entailing that there are no genuine (no factual) autonomous non-basic sciences.

of translation manuals, for example, facts about human perception, desire, and reasoning (1998, 158).

Hookway stresses that Quine's restriction of the relevant evidence is not unmotivated. Quine regards ordinary psychology talk as not strictly factual, and holds that for scientific purposes it should yield to a suitable regimentation which will be behavioristic or physiological in character (1998, 160-62). But, I submit, though it is open to Quine to appeal to the non-factuality of mentalist psychology in justifying his limitation of evidence for the indeterminacy of translation to the behavioral, the problem this poses for his attempt to draw consequences of that indeterminacy for the mind is readily apparent. Quine wishes to argue from the indeterminacy of translation to the factual illegitimacy of intentional psychology. But his reliance on exclusively behavioristic evidence and repudiation of such evidence as could be afforded by intentional psychology can be justified only by the prior rejection of intentional psychology. That is, in attempting to infer the non-factuality of intentional psychology from the indeterminacy of translation Quine simply begs the question in a rather crass fashion.

A Rejoinder Foreclosed

Precisely at this point a rejoinder that would be open to a thoroughgoing interpretationist such as Davidson seems closed off to Quine. In the case of Davidson, it is clear that he intends to give an interpretationist account of both meaning and psychology. In fact, he insists that this requires giving a unitary account wherein facts of meaning and of psychology are attributed all at once in an interconnected fashion. On such a view, there would be clear motivation for bracketing ordinary-psychological evidence from figuring into one's account of radical translation. For inasmuch as such an interpretationism proposes to be a constitutive account of psychology as well as meaning, it cannot advert to psychological evidence without circularity. The restriction to purely behavioristic evidence that Quine insists on, then, would be justified by its allowing one to avoid such circularity.[25] But it is far from clear that this maneuver is available to Quine. For there seems no definitive indication that Quine is an interpretionist with respect to psychology. At least on its face, Quine's account of radical translation is an account of meaning alone, and, unlike Davidson's account, is not obviously embedded in a general account of intentionality. Nor does Quine give anything that could be construed as a separate interpretationist account of mental states. Granted, Quine shares Dennett's instrumentalist outlook with respect to the mind—and Dennett is an interpretationist with respect to that domain—but instrumentalism does not appear to entail interpretationism with respect to psychology.

Quine as Hermeneuticist

Perhaps enough has been said to give color to Quine's general view of meaning and psychology. Before turning to the role of charity in Quine, however, I would like to briefly comment on the extent to which Quine's view of the meaning and mind is aptly characterized as hermeneutical. The issue is in what degree for Quine these domains float free of natural-scientific methodology and sources of evidence. As the quote from *Word and Object* above indicates (see p. 22), Quine diverges from Brentano in denying the importance of an "autonomous science of intention." His view seems to be that mind and language are not properly scientific

[25] But to exclude other sorts of non-intentional evidence, especially neurophysiological evidence, as do Quine and Davidson, requires some other justification.

domains at all. Moreover, in the case of the latter, his restriction of the sources of evidence for radical translation (essentially, to stimulus-meaning) runs counter to the confirmation holism which Quine—rightly—takes to govern methodology in the sciences. With respect to psychology, Quine is less explicit about what methods he takes to govern the attribution of intentional states. But if Quine is a normativist (see the next section), and if Davidson is correct in arguing that normative constraints on the propositional attitudes entail that there are no strict psychophysical laws (see p. 13 above), then his normativism would itself seem to commit him to the inapplicability of natural-scientific methods to intentional states. So there appears ample reason to regard Quine's view of psychology (and language) as hermeneutical or insularist.

Quinean Charity

In the present section, I shall rough out a portrait of Quinean charity. But there exist competing interpretations of Quine's charity doctrine and the role which it plays within his views of language and mind. So I shall first set out what appear to be the main interpretative options. Then I shall examine some of the chief sources for Quine's charity doctrine in his writings with a view towards adjudicating among the alternative interpretations as well as pinning down other matters of interpretation on which the major readings are silent. My conclusions in this regard will be somewhat tentative, but perhaps I can excuse this by noting that my consideration of Quinean charity is largely meant to prepare for and illuminate the sorts of issues that will arise in considering, in Ch. 2, the charity doctrine of latter-day normativists such as Davidson.

The Pragmatic Interpetation

Christopher Hookway develops a pragmatic interpretation of Quinean charity. As mentioned above, Quine holds that radical translation is indeterminate: the sum of evidence relevant to the translation of a native's language into an interpreter's language substantively underdetermines how it is to be translated. But Quine appears to acknowledge a category of "supplementary canons" which, he suggests, linguists use to narrow the range of possible translations of words and phrases. So, for example, they will translate 'gavagai' as 'rabbit' rather than 'rabbit part', Quine states, because they will assume that "the more conspicuously segregated wholes are likelier to bear the simpler terms" (1960, 74). Hookway interprets Quine here as acknowledging a class of purely pragmatic criteria which, though their employment is justifiable by their utility, do not in any way support the truth of the analytical hypotheses they favor (1988, 135).

Moreover, Hookway reckons Quine's principle of charity to this class. We prefer translations that maximize our agreement with others, so "the best translation will be one that minimizes inexplicable error". This "makes it easier for us to learn from their testimony, and helps us to co-operate with them" (1988, 136). But, again, such a principle does not diminish indeterminacy. It is merely a useful device for living with that indeterminacy.

The Constitutive Interpretation

Hookway contrasts Quine with Davidson in this regard. Whereas for Quine charity is a mere pragmatic expedient, for Davidson it is "constitutive of correctness." An adequate translation must show that natives' beliefs are mostly true (1988, 170). Hookway rather offhandedly mentions that Quine might hold that we are compelled to translate natives in a way that sees them as obeying our logical principles, but suggests that "his views on this matter are not clear" (1988, 136). So, by and large, Hookway minimizes any role of charity as a constitutive constraint on translation for Quine. Moreover, if he is right, then at bottom charity is not a constitutive constraint on psychology either. For if it were, it would perforce constrain translation as well.

A constitutive reading of Quinean charity, however, cannot be casually dismissed. As Ed Stein points out, Quine's discussion of "prelogical mentality" (1960, 58) gives the impression that "our logic should be *imposed* on the people we translate" (Stein 1996, 113). Stein observes that this might suggest that Quine adheres to a "strong principle of charity" according to which "people should *always* be interpreted as rational" (1996, 24). Such a principle would, in effect, make (theoretical) rationality a necessary condition for the possession of mentality or language; that is, it would render charity a constitutive constraint on psychology and language.

The Heuristic (or 'Weak') Interpretation

Stein notes, however, that there is also textual evidence that seems to undercut such a strong reading of Quinean charity (1996, 113). Quine's statement that "one's interlocutor's silliness, beyond a certain point, is less likely than bad translation" (1960, 59) strongly suggests at least the possibility of an agent's illogicality, which runs counter to Stein's strong reading of charity. Accordingly, Stein mentions an alternative "weak version" of Quine's principle of charity on which charity is "defeasible (it says that people should be interpreted as rational *unless* there is strong evidence to suggest otherwise)" (1996, 24).

Hans-Johann Glock explicitly advocates a reading of Quinean charity more or less along these lines.[26] He states that, for Quine, charity is merely a “heuristic” device that “enhances the prospects of interpretation” (Glock 2003, 184; cf. 238). What Glock seems to mean is that our interpretations are more likely to be correct to the extent that they are charitable. Thus, he appears to attribute Stein’s weak charity to Quine. He maintains, further, that in contrast to Davidson, “charity” for Quine “is not constitutive of the concepts of translation or interpretation” (2003, 238; cf. 33). So the heuristic/weak reading of charity represents a second non-constitutive reading of Quinean charity, alongside Hookway’s pragmatic reading.

The Interpretations Compared

It is important, however, not to be misled by the terms in which the interpretative controversy surrounding Quinean charity is cast. The contrast between the terms ‘constitutive’ and ‘pragmatic’, in particular, may suggest that the interpretative issue somehow turns on whether Quine views language and mind as factual or non-factual domains. But, in truth, the major proponents of each of the three interpretative options described—constitutive, pragmatic, and heuristic—take for granted that Quine regards meaning and mind as non-factual. Rather, the bone of contention among the readings, in the first instance, is simply whether for Quine charity is an obligatory demand on interpretation. This will be so if on Quine’s account interpretations cannot even be generated unless charity is applied. The constitutive reading takes this to be the case, whereas the pragmatic and heuristic readings do not. The latter two readings hold instead that Quine’s account generates

[26] Apparently, he does so on the basis of the textual evidence just discussed. Cf., e.g., (Glock 2003, 172 and 184).

interpretations independently of charity. On the pragmatic reading, charity enters in only as a device for winnowing down in useful ways the range of acceptable, independently generated interpretations. On the heuristic reading, it merely serves to increase an interpreter's likelihood of correctly latching onto one of the range of acceptable, independently generated interpretations. Both pragmatic and heuristic readings, then, agree in assigning charity a somewhat marginal place in Quine's account of meaning and mind.

There are important differences between the heuristic and pragmatic readings, however. A salient difference between the two is perhaps best approached by considering the implications of Quine's famous rejection of the traditional distinction between analytic and synthetic truths (1980). Once this distinction is rejected, so is the—in some measure—correlative distinction between *a posteriori* and *a priori* knowledge. Perhaps it is apt to say that for Quine all truths are synthetic and all genuine knowledge *a posteriori*.

Now the heuristic reading seems to accord charity an *a posteriori* status. The heuristic interpretation of charity seems to regard it as an empirical truth that agents generally tend to be logical, etc.[27] Glock at least seems to see to the matter so, for he cites "Quine's naturalism" as a reason that Quine cannot share Davidson's constitutive approach to charity (2003, 33).[28] The heuristic interpretation asserts a presumption of logicality, etc.: one shouldn't attribute obvious error without

[27] Some qualification is in order here: In light of Quine's ultimate denial of factuality to the intentional idiom, the heuristic reading cannot strictly be said to regard it as an aposteriori or empirical *truth* that agents tend to be logical, etc. Nonetheless, it regards the provenance of the principle of charity as empirical to the extent that it is based on something like experiential 'evidence'.

[28] Whether Glock is correct in thinking that Quine's naturalism rules out his taking a constitutive view of charity is a point that I address below.

(empirical) evidence that overrides that presumption. Presumably, that is because there is taken to be overwhelming (empirical) evidence that people are generally logical, etc. Things stand quite differently with respect to the pragmatic reading. In contrast to the heuristic reading, the pragmatic reading is decidedly non-empirical. Nonetheless, it might be misleading to characterize the principle of charity on this reading as *a priori*. Charity on this reading is seen as a purely optional, instrumental expedient. So the notion of the *a priori*, which is bound up with that of necessity, does not get a foothold.[29]

Both the heuristic and pragmatic readings, though, raise the possibility that charity for Quine is not a constitutive constraint on intentional attitudes, which is significant for two reasons. First, if either interpretation is correct, then the status Quine assigns the principle of charity differs markedly from that assigned it by latter-day proponents of charity such as Davidson and Dennett, who clearly assign it a constitutive role.[30] Moreover, if charity *is* non-constitutive for Quine, this raises an issue whether Quine is even genuinely to be classed as a normativist.

Usage of the phrase ‘principle of charity’ suggests that it is neutral with respect to its status as *a posteriori* or *a priori*, constitutive or non-constitutive[31]. However, usage of the term ‘normativism’—such as it is—seems to reserve it for

[29] If charity is constitutive for Quine, by contrast, there is still, however, some problem in describing it as *a priori*. The problem is that Quine *ultimately* denies factuality to the intentional idiom and thereby renders notions of truth and knowledge inapplicable to it. Cf. above p. 30, n. 27. However, I take the necessary, non-empirical character of charity on the constitutive reading to justify the use, in a rough-and-ready way, of the designation ‘*a priori*’.

[30] Actually, the situation with respect to Dennett is a bit complex. Dennett seems to be an instrumentalist with respect to the mind (cf. Rey [1994]), which may initially suggest the pragmatic reading of charity. But the principle is mandatory with respect to psychological attribution in a way it would not be for Quine. So charity, I think, is best seen as constitutive for Dennett.

[31] The usage of Hookway, Stein, and Glock with respect to the interpretation of Quinean charity which I have discussed above demonstrates as much.

views which regard charity as an *a priori* constitutive constraint for the possession of propositional attitudes. In any case, that is how I am using the term in this dissertation. This provides me with a ready designation for the view which at which I am leveling in this project. For it is no part of my plan to argue against charity *per se*, where there is no implication that the principle is accorded *a priori* constitutive status. It is the view that sees rationality as an *a priori* condition on possession of propositional attitudes which I take as my specific target. I am concessive with respect to the possibility that the best account of propositional content may turn out to place normative constraints on the possession of (at least some) concepts. Such would be the case if an inferential-role or so-called ‘two-factor’ theory of content proved to be the correct one.[32] But the truth of such a theory, I am inclined to believe, would be one that could be discovered only *a posteriori*.[33] At all events, in my usage, such theories will not be reckoned ‘normativist’.

With the term’s usage fixed as I have done, on either the heuristic or pragmatic reading, Quine decidedly does not count as a normativist, since either reading renders Quinean charity non-constitutive. One issue, then, as I proceed to examine the chief sources in Quine’s writing for his views on charity, will be whether the bulk of the evidence supports a constitutive (normativist) or non-constitutive reading. At stake is the extent of continuity between Quine and his clearly normativist successors.

[32] See, e.g., Rey (1997, 237-63) for characterizations of theories of these types.

[33] Of course, one possible outcome is that despite the existence of *a posteriori* normative constraints on propositional attitudes, they could fall far short of the rather severe ones which normativists typically impose. I return to this matter in Ch. 2.

Quine’s Pronouncements Concerning Charity

One of the *loci classici* for Quine’s charity doctrine is his *Word and Object*, sect. 13, “Translating Logical Connectives” (1960, 57-61). In the context of his account of radical translation, Quine here proposes a method for identifying and translating those locutions of a language that express truth functions. Quine states “semantic criteria” for various truth functions in terms of natives’ assent or dissent to sentences. So, for example, “The semantic criterion for negation,” Quine writes, “is that it turns any short sentence to which one will assent into a sentence from which one will dissent, and vice versa” (1960, 57). Again, the criterion for conjunction is that an expression turns two short sentences into a sentence to which one will assent just in case one will assent to both of its component-sentences. The restriction to short sentences is meant to rule out the possibility of a native’s diverging from criteria simply as a result of the confusion which, Quine thinks, long sentences can engender. Once an expression is identified as expressing a truth function in this way it may be straightforwardly translated with the corresponding English expression (e.g., ‘not’ for negation).

Quine observes that his approach to the translation of connectives conflicts with the doctrine of “prelogical mentality,” which he attributes to the anthropologist Lucien Levy-Bruhl (1960, 58n1). Ignoring niceties of Levy-Bruhl’s original doctrine, Quine focuses on a “caricature” of the view, according to which “there are pre-logical peoples who accept certain simple self-contradictions as true” (1976, 109), that is, “sentences translatable in the form ‘*p* and not *p*’” (1960, 58). Quine rightly notes that his views about the translation of connectives render this hypothesis “absurd.” For by the criterion for conjunction, a native will assent to '*p* and not *p*'[34] just in case they will both assent to '*p*' and assent to 'not *p*'. But by the criterion for negation, they will assent to 'not *p*' just in case they will *dissent* to '*p*'. That is, the supposition that a native accepts a sentence translatable as '*p* and not *p*' entails the absurdity that they both assent to and dissent to '*p*' and, therefore, must be rejected. Accordingly, on Quine's account, pre-logical mentality is impossible. Prelogicality is ruled out because, Quine maintains, "better translation imposes our logic" on natives: "fair translation preserves logical laws" (1960, 58-59). So much so, Quine holds, that even when someone seems to espouse a logic in which a logical law like non-contradiction is rejected, we are led to reinterpret their English statements rather than attribute to them a contradictory logic (though this means overriding the usual homophonic translations).

Such a practice rests, Quine asserts, on "The maxim of translation . . . that assertions startlingly false on the face of them are likely to turn on hidden differences of language." He notes further, "The common sense behind the maxim is that one's interlocutor's silliness, beyond a certain point, is less likely than bad translation . . ." (1960, 59); and in a footnote, he suggests an affinity between this idea and N. L. Wilson's "principle of charity" (1960, 59n2). In effect, then, Quine here states his own principle of charity.[35]

[34] Or, more precisely, the corresponding native sentence (similarly for '*p*' and for 'not *p*'). Also, it is assumed here that sentence '*p*' is sufficiently short.

[35] Quine does not at this point refer to the principle which he enunciates here as a 'principle of charity'. But a bit later (1960, 69n1), he identifies the principle of charity as having been a focus of sect. 13. There is also the use of the phrase 'fair translation' (1960, 59), 'fair', of course, being a synonym of 'charitable'.

Aside from the appeal Quine makes to charity to ground his views about the role of logic in translation, charity seems to enter Quine's account of translation in *Word and Object* at a second point. As I noted above (p. 17), Quine identifies as a constraint on translation that a manual should map sentences which are stimulus-analytic (-contradictory) for natives onto English sentences that are so for English speakers. Quine remarks, however (1960, 69), that this injunction is not to be taken altogether strictly. He says that a manual may permit a few native sentences which are stimulus-analytic to be translated by English ones that are not if the manual has the merit of being markedly simpler than alternative manuals. Nonetheless, stimulus-analyticity is generally to be preserved in translation, and Quine underwrites this by an appeal to charity. He writes, "the more absurd or exotic the beliefs imputed to a people, the more suspicious we are entitled to be of the translations" (1960, 69); and in this connection he refers back to his earlier discussion of charity in sect. 13.

Such are Quine's pronouncements about charity in *Word and Object*. Leaving aside Quine's remarks about stimulus-analyticity for the moment and focusing on sect. 13, there are essentially four elements in Quine's account of charity requiring coordination: (1) Quine's semantic criteria for the translation of truth-functional connectives, (2) his claim that "fair translation preserves logical laws" (1960, 59), (3) the phenomena such as prelogicality and deviant logic that Quine wishes to rule out, and (4) the bedrock principles, namely, the "maxim" that assertions startlingly false on the face of them are likely to turn on hidden differences of language" and "the common sense behind the maxim . . . that one's interlocutor's silliness, beyond a certain point, is less likely than bad translation . . ." (1960, 59).

A few comments about the relations among these elements are in order. First, although Quine appeals to (1) the semantic criteria for the truth-functional connectives to rule out natives' accepting a sentence of the form '*p* and not *p*' (cf. 3), these criteria do not rule out illogical phenomena on the part of natives in general. They do not rule out natives' acceptance of sentences (e.g., of the form 'all *P* are not *P*') whose logical falsehood is not a matter of their truth-functional structure.[36] Rather, it is (2) Quine's insistence that translation preserve logical laws that is responsible for ruling out the illogical phenomena. Further, it is (4) the bedrock principles that for Quine are supposed to provide the ultimate grounding for his view about logic encapsulated in (2).

The Scope of Quinean Charity

This, however, raises a slight interpretative difficulty. To see how this is so, it will be useful to set out a distinction, owing to Donald Davidson, between two sorts of charity principles, namely, Principles of *Correspondence* and Principles of *Coherence*. Roughly, whereas the former concern the truth of an agent's statements or beliefs, the latter concern the degree of rational consistency among an agent's statements, beliefs, or other propositional attitudes (cf. Glock 2003, 194-95). An issue with respect to the interpretation of Quinean charity, then, is whether he advances a Principle of Coherence in addition to a Principle of Correspondence.

Now the bedrock principles (4), whatever their import in detail, seem to preclude an agent's making "startingly false" assertions or having beliefs of that sort. So they pretty clearly amount to a Principle of Correspondence. Things are less

[36] Moreover, Quine emphasizes that semantic criteria comparable to those for the truth-functional connectives cannot be formulated for the non-truth-functional elements of language such as quantificational expressions. See (1960, 60).

straightforward, however, with respect to Coherence. Quine's insistence (2) that translation preserve logical laws may suggest a Principle of Coherence inasmuch as logic formulates principles which bear on the consistency of sets of propositions and, therefore, beliefs. Indeed, any valid sequent *p*, *q*/*r* is tantamount to a consistency constraint on beliefs, namely, one requiring that one not simultaneously entertain the beliefs *p*, *q*, and not *r*. But Quine's examples of translations which are unacceptable because they would violate the injunction to preserve logical laws all involve the attribution of single contradictory beliefs, not inconsistent sets of beliefs (e.g., the belief that *p* and not *p*).[37] In fact, the "logical laws" Quine seems to have in mind in the first instance are the "tautologies" and other "logical truths" (1960, 60). Moreover, Correspondence Principles like the bedrock principles of (4) imply nothing about *coherence* among propositional attitudes.

But Quine's semantic criteria (1) for translating the truth-functional connectives entail a degree of consistency among an agent's beliefs. For example, the criterion for negation rules out an agent's simultaneously believing *p* and not *p*. In fact, Quine's semantic criteria rule out an agent's having any inconsistent set of beliefs whose inconsistency is a matter of the sentences' truth-functional structure. But, of course, there is a very close relation between logical falsehood and inconsistency: a set of sentences *p*, *q*, *r* is inconsistent just in case their conjunction *p*&*q*&*r* is a logical falsehood. In light of this fact, then, Quine's account of charity seems to preclude an agent's having at least *some* inconsistent sets of beliefs whose inconsistency is *not* a matter of truth-functional structure. For if there is any such set

[37] This statement requires the obvious qualification that a single contradictory belief itself virtually constitutes an inconsistent set of beliefs, namely, the singleton which has that belief as its only member.

p, *q*, *r* whose corresponding conjunction is “startingly false” for an agent, then by (4) the bedrock principles it apparently should not be ascribed to the agent. But then it follows from Quine’s semantic criterion for conjunction that the agent does not hold the inconsistent set of beliefs *p*, *q*, *r* either. The upshot, then, is that even though Quine’s account seems to emphasize Correspondence charity, nonetheless, it includes elements of Coherence charity as well.

Nonetheless, there are fairly severe limits to the scope it grants to Coherence, that is, rationality-charity. For, since Quine’s account of charity applies only to beliefs, its scope is restricted to *theoretical* rationality. The whole domain of practical rationality, inasmuch as it concerns desires, intentions, and actions in addition to beliefs, is excluded from its sphere of application.[38] But even within theoretical rationality, its scope is quite narrow. This can be seen by considering its significance for the part of theoretical rationality that is *procedural*, which concerns the processes one follows in forming beliefs, as opposed to the part that is *statal*, which concerns the relations among the products of such processes. The applicability of Quinean charity to processes of belief-formation is limited. It restricts inferences whose invalidity would introduce a logical inconsistency into an agent’s belief set. But it is less clear that it restricts inferences whose conclusions, though not entailed by the premises of the inference, introduce no such inconsistency. So many—in fact, most—deductively invalid inferences slip by Quine’s charity principles. Indeed, they fail to constrain most non-deductive (i.e., inductive and abductive) inferences for the same reason. Quine’s charity principles also fail to constrain cases of procedural

[38] As is any other sort of rationality—such as emotional rationality—that does not exclusively concern beliefs (though on a cognitive theory of emotion, where emotions are understood as kinds of beliefs, emotional rationality reduces to a sort of theoretical rationality).

theoretical irrationality that are not a matter of logical inference at all, such as cases of self-deception. The scope of Coherence charity in Quine, then, is quite narrow.[39]

Other Interpretative Issues

Issues of scope aside, other points with respect to the interpretation of the content of Quinean charity demand consideration. First, there is an issue of its strength even within its sphere of application. Now, on their face, the bedrock principles suggest a certain moderation in Quinean charity. The impression they impart is that it is only "*startingly* false" beliefs that one needs to be wary of ascribing to agents, that the epistemic impropriety of holding false beliefs admits of degrees, and only "beyond a certain point", only when the error involved in holding a false belief would rise to a certain level of egregiousness, must one avoid attributing it. This is quite clear, moreover, from Quine's formulation of charity in his *Philosophy of Logic* as the principle to "'Save the obvious'" (1986, 82): it is only obviously false beliefs that one must not ascribe to an agent.

But, at least with respect to logic, this impression of moderation is belied by other statements that Quine makes in the same place (1986, 82-83). He gives the impression that logic constrains translation quite strictly: "If a native is prepared to assent to some compound sentence but not to a constituent, this is a reason not to construe the construction as conjunction." In effect, "we build logic into our manual of translation" (1986, 82). Moreover, Quine states that "every logical truth is obvious, actually or potentially." Consequently, "The canon 'Save the obvious' bans any manual of translation that would represent the foreigners as contradicting our

[39] Issues of the scope of various charity principles will be something of a *leitmotif* in subsequent chapters, for aside from purely interpretative questions, such issues have a bearing on the success of my arguments against charity constraints.

logic" (1986, 82-83). So Quine evidently holds that charity with respect to logic is ideal: it rules out all logical falsehoods (and inconsistency)[40].

Christopher Cherniak, however, observes that Quine's view here is quite problematic. "Such a translation principle," Cherniak writes, "excludes an agent from accepting even the most obscure inconsistencies," and "implies triviality of significant portions of the deductive sciences" (Cherniak 1986, 96). Quine writes, "every logical truth is obvious, actually or potentially. Each, that is to say, is either obvious as it stands or can be reached from obvious truths by a sequence of individually obvious steps" (Quine 1986, 82-83). So, as Cherniak notes, Quine here distinguishes between the actual obviousness which attaches to the axioms of a logistic from the mere potential obvious which attaches to the theorems of the system. But in holding that logical falsehoods must not be ascribed to agents, Quine illicitly (and implausibly) treats all logical truths as if they were actually obvious.

Finding such ideal charity with respect to logic far-fetched, Cherniak prefers to consider the implications of a more moderate principle: "'Better translation favors the subject's not accepting the more obvious inconsistencies'" (1986, 96). But he stops short of attributing this principle to Quine. In fact, he earlier writes that "Quine's translation methodology . . . presupposes an ideal consistency condition"

[40] A qualification, however, may immediately suggest itself. Quine writes that "corrigible confusions in complex sentences" are possibly exempted from charity constraints on logic (1986, 83). Cf. his formulations of the semantic criteria for the truth-functional connectives in *Word and Object*, which are qualified so as to apply only to short sentences (1960, 57-58). This may seem to greatly limit the strength of Quinean logic-charity. But such restrictions based on sentence-length seem to be allowances for failures to *parse* long sentences, that is, for confusions about the *meanings* of sentences, rather than for an agent's actually entertaining a logical falsehood (or inconsistent beliefs). Thus, e.g., a native may assent to a long conjunction but dissent to one of its shorter constituents when queried simply because they are (temporarily) confused about the meaning of the conjunction as a whole. Such linguistic confusion does not constitute epistemic or rational error.

(1986, 18). So Cherniak apparently attributes a very strong charity to principle to Quine as far as logic is concerned. I follow Cherniak in this assessment and shall proceed on the assumption that Quinean logic-charity is ideal.

That Quine regards logic-charity as ideal makes the question what demarcation ‘obvious falsehood’ represents for Quine moot with respect to this sort of charity. However, Quine’s principle to ‘Save the obvious’ takes in more than logic. Quine writes, “Being thus built into translation is not an exclusive trait of logic. If the natives are not prepared to assent to a certain sentence in the rain, then . . . we have reason not to translate the sentence as ‘It is raining’” (1986, 82). His point is that to render the sentence as ‘It is raining’ would mean attributing to them what in the circumstances is an obviously false belief and, therefore, violating his charity principle. But, of course, not all false beliefs are in any intuitive sense obviously so. As Quine puts it, “the incidence of obviousness” in most domains is less than that of logic. So application of Quine’s maxim beyond logic requires clarity as to what constitutes obvious falsity.

Quine, in line with the general behavioristic orientation of his account of translation (cf. p. 20 above), insists that he is “using the word ‘obvious’ in an ordinary behavioral sense, with no epistemological overtones.” The standard that he applies is based on the pattern of assent queried sentences receive: “When I call ‘1 + 1 = 2’ obvious to a community I mean only that everyone, nearly enough, will unhesitatingly assent to it . . .; and when I call ‘It is raining’ obvious in particular circumstances I mean that everyone will assent to it in those circumstances” (1986, 82).[41] Presumably, then, obvious *falsity* for a community is to be assessed similarly, except on the basis of universal dissent rather assent. As Cherniak notes, however, there is a question whether it is with reference to the translator or to the natives that obviousness is to be judged for purposes of charity (1986, 96). Cherniak observes, further, that Quine's discussion of charity in *Word and Object* suggests the former. In fact, the criterion of obvious falsity just mentioned could not sensibly be employed in a charity principle couched in terms of obvious falsity for natives.[42] So the operative standard for Quine appears to be obvious falsity for the translator's community.[43] In interpreting speakers, the translator is to avoid ascribing beliefs which are obviously false in his own community.

Holistic and Non-Holistic Charity

In further characterizing Quinean charity, a useful distinction can be made between holistic and non-holistic versions of charity (cf. Stein 1996, 124-27). Whereas non-holistic versions constrain beliefs, inferences, actions, etc., individually, holistic versions constrain an agent's whole system of beliefs, etc. Thus, for example, a principle to attribute to an agent only rational inferences counts as non-holistic,

[41] Quine formulates this criterion of obviousness in the context of stating, as a constraint on translation, the principle that in rendering a language we should "make the obvious sentences go over into English sentences that are true and, preferably, also obvious" (1986, 82). Essentially, this is the principle, discussed above (p. 17) that stimulus-analytic sentences should be rendered with stimulus-analytic sentences, broadened to include sentences where there is universal assent relative to particular circumstances. Presumably, Quine takes this definition of 'obvious' to apply to its use in his charity principles as well.

[42] The problem is that, to take the simplest case, the principle would enjoin one not to attribute to a native beliefs that all natives dissent from. But any such belief is one that the native in question also dissents from! That fact alone should deter one from attributing the belief to the native. So charity would be empty. This problem is avoided when the standard is taken to be obviousness for the translator's community.

[43] Cherniak notes that such a principle is implausible: "we cannot egocentrically assume that what is startingly false for the observer must be startlingly false for the subject" (1986, 96-97). Therefore, Cherniak concludes that charity should be couched in terms of obviousness for natives. But Cherniak stops short of maintaining that such is Quine's intention, and, in fact, the textual evidence seems to support the reading of obviousness in terms of the translator's community.

since it amounts to a test which individual inferences can be judged to pass or fail without reference to the (ir)rationality of other inferences. A principle, by contrast, to ensure that one attribute a preponderance of rational inferences to an agent is plainly holistic, since it constitutes a test applying to an agent's inferences not individually but *en masse*.

I think it can be safely concluded that Quine's charity principle is non-holistic in the sense defined. For Quine's injunction not to attribute obviously false beliefs to an agent functions as a filter on beliefs individually. By way of comparison, a principle that tolerates some obviously false beliefs provided, say, they represent a small proportion of an agent's overall system of beliefs would be clearly holistic—but pretty clearly not Quinean. Granted, Quine's statement that "one's interlocutor's silliness, beyond a certain point, is less likely than bad translation" (1960, 59) might seem to permit a holistic construal where the "certain point" is taken to be a matter of a tipping point of the aggregate "silliness" (i.e., obvious falsity) of an agent's ascribed system of beliefs. But Quine nowhere enunciates this thought, and when he discusses charity his focus always appears to be on the (un-)acceptability of individual ascriptions, uncompromisingly ruling them out when they are what he takes to be obvious falsehoods.

Wilson's Principle of Charity

Other facets of Quinean charity can, I think, be illuminated by a comparison with the principle of charity enunciated by N. L. Wilson (Wilson 1959). Moreover, such a comparison seems in order inasmuch as Quine himself suggests that an affinity exists between his and Wilson's versions of charity (1960, 59n2). Wilson enunciates a principle of charity in the context of a consideration of the question “How does a name in use get its significance?” (1959, 529), that is, in the context of formulating a theory of the reference of proper names in an individual’s idiolect.[44] On Wilson’s account, we determine the reference of, say, the name ‘Julius Caesar’ in Charles’ language (to use Wilson’s example) by examining the statements he asserts containing the name. More particularly, we apply what Wilson dubs “the Principle of Charity” which enjoins us to “select as designatum that individual which will make the largest number of Charles’ statements [containing the name] true” (1959, 532).

As a theory of reference, Wilson’s account appears a version of Fregean descriptive theory since, in effect, it makes reference a matter of an item’s satisfaction of predicates, that is, of descriptions being true of it. It has particular affinities, however, with John Searle’s roughly contemporaneous cluster theory of names (Searle 1958) inasmuch as it does *not* make reference depend on the satisfaction of some single predicate (or conjunction of predicates). It differs from it in some particulars, however. First, Searle’s formulation is more social, concerned with the reference of names in a community.[45] Also, Wilson’s formulation is comparative whereas Searle’s is not. Whereas for Wilson, the referent of a name is whatever satisfies more of the predicates asserted of a name than other items, for Searle the

[44] With respect to the issue of priority in formulating a charity principle: In (1976, 109), Quine expresses views about the translation of logical particles and the impossibility of prelogicality which amount to implicit charity constraints. Moreover, he states “there can be . . . no stronger evidence of bad translation than that it translates earnest affirmations into obvious falsehoods” (1976, 113), which is extremely close to his reasonably explicit statement of charity in *Word and Object* (1960, 59). So perhaps credit should be accorded Wilson merely for coining the *term* ‘principle of charity’.

[45] The version of the cluster theory which Saul Kripke formulates in *Naming and Necessity* for critical purposes, however, resembles Wilson’s theory in being formulated in terms of an individual’s idiolect (cf. Kripke 1980, 71).

referent is whatever, if anything, satisfies most of them. But, nonetheless, their theories are largely of a piece.

Wilson's account, further, does in fact impose a charity constraint, though of a distinctive sort, and only within a somewhat circumscribed sphere. As a principle of interpretation, it has the effect of maximizing the truthfulness of an individual's beliefs such as are expressed in assertions involving proper names. Maximization principles like this, which enjoin that one maximize an agent's truthfulness or rationality, represent an important species of charity principle. They are inherently holistic in nature, constraining not the ascription of individual propositional attitudes but rather whole sets of them at once.[46] Moreover, they are inherently comparative in the sense that whether an interpretation meets the constraint they lay down cannot be determined by considering the interpretation in isolation but, rather, only by considering how it stacks up against other candidate interpretations.

As mentioned, Quine cites Wilson's principle at the point that he enunciates his bedrock principles. So he clearly sees an affinity between his own charity principle and Wilson's. But their similarity seems confined to the fact that they are both Correspondence principles, that is, principles constraining the truthfulness of an agent's beliefs. For Wilson's charity principle is maximizing and, therefore, holistic

[46] In Wilson's case, the holism is quite moderate, since it is only beliefs (indeed, only a subset of them) that are jointly constrained by his charity principle in any particular application of it.
It merits noting, however, that though Wilson's principle is holistic in its implications for charity, it is non-holistic with respect to semantics. For it states conditions which determine the reference of proper names in isolation from the question how other parts of speech gain their reference. Indeed, so pronounced is Wilson's semantic non-holism that he is seduced into a vicious circularity: his account of the reference of individual proper names takes for granted that other proper names have already acquired a reference (cf. 1959, 530: "Let us suppose . . . that we know the significance which Charles attaches to expressions other than 'Caesar' . . . .).

and comparative, whereas Quine's is none of these things.[47] Granted, as I mentioned, it is not impossible to construe Quine's statement of the bedrock principles as holistic. But even then his principle would differ significantly from Wilson's in character. For, on a holistic reading, Quine's statement that "one's interlocutor's silliness, beyond a certain point, is less likely than bad translation" (1960, 59) would amount to the claim that the number or proportion of obviously false beliefs in an agent's belief-set cannot surpass a certain threshold. Threshold Principles like this one represent an important type of charity principle.[48] But they lack the maximizing, comparative character of Wilson's principle. So even on the—implausible—holistic reading Quine's principle would differ markedly from Wilson's.

The Cultural Historian's Principle of Charity

It is instructive, further, to compare Quine's (and Wilson's) principles of charity to the less technical notions of charity often appealed to by intellectual historians (such as historians of philosophy) as justification for their interpretative practice. First, the scope of the latter notions seems to be narrower in one respect, for these less technical charity principles are typically intended to apply only to people, such as famous philosophers, who can safely be presumed to be smart! The intellectual historian's charity is in this regard markedly undemocratic. Second, though charity of this sort resembles Quine's non-holistic charity in that it typically aims to avoid ascribing obvious falsehoods (egregiously invalid inferences, etc.), this

[47] Moreover, even though both principles ostensibly constrain the truthfulness of beliefs, in fact Quine's principle concerns the degree of acceptance of beliefs in a native's community. Wilson's principle, by contrast, seems to concern the literal truthfulness of an individual's assertions involving a proper name.

[48] Of course, Threshold Principles can concern rationality as well as truth—a fact that will loom large in considering Davidson's normativism in succeeding chapters.

is typically tempered by a readiness to ascribe some such lapses if the interpretation that yields them at least—among available interpretations—maximizes the overall truth, coherence, etc. of an author's claims. So at bottom such charity—in contrast to Quine's (and like Wilson's)—seems to have a holistic, maximizing character.[49] Last, it is unlikely that the wielders of such principles regard them as possessing anything other than a heuristic status: greater truth-preservation, coherence, etc., constitute a presumption in favor of an interpretation, but such a presumption can be overcome.[50]

Adjudicating Among the Interpretations of Quinean Charity

It is time now to consider what status Quine assigns his principle of charity—pragmatic, heuristic, or constitutive. Of the three interpretations, the claims of the pragmatic reading seem most precarious. As Hookway points out, Quine does seem to acknowledge a category of purely pragmatic "supplementary canons" for the construction of translation manuals whose employment is justified solely by their utility (Quine 1960, 74). But there seems no textual evidence to support Hookway's view that Quine's principle of charity should be included among them. Quine makes no mention of charity in discussing them. Moreover, Quine's mention of the "supplementary canons" is relegated to one rather short paragraph of *Word and Object*. His remarks concerning them seem to amount to little more than aside. By contrast, Quine's account of charity (1960, sect. 13, 57-61) is to all appearances a

[49] That such charity is limited in its application to smart people lends it a greater *prima facie* plausibility than broader charity principles like Quine's. For the smarter someone is, the greater the likelihood that the truth- and rationality-maximizing interpretation will apply to them. But given the limits on human intellect, a principle which requires maximization come hell or high water would possibly be too strong.
I postpone discussion of a further element of such charity till Ch. 2, namely, that the principles that intellectual historians actually employ are closer to what Richard Grandy dubs 'the principle of humanity' (Grandy 1973) than to, say, Quinean or Davidsonian charity principles.

[50] This attitude, in fact, suggests the wrongheadedness of properly normativist views of charity which regard it as a constitutive constraint.

prominent and substantial element of his general account of radical translation. Hookways's reading, then, significantly mislays the emphasis in Quine's account, assigning the supplementary canons an import out of proportion to their true significance.

Hookway is abetted in this by a further peculiarity of his reading. Hookway clearly regards the crux of Quinean charity as his injunction that sentences that are stimulus-analytic (or stimulus-contradictory) for natives are to be rendered with sentences that are so for English-speakers (Quine 1960, 68). Hookway altogether downplays Quinean logic-charity. He writes, "although Quine *may* believe that we are constrained to read our Logic into the verbal behaviour of the natives, his views on this matter are not clear" (Hookway 1988, 136). In fact, there is little clearer in Quine's account of charity than that translation imposes our logic: "The canon 'Save the obvious' bans any manual of translation that would represent the foreigners as contradicting our logic" (Quine 1986, 83). To all appearances, logic-charity forms a central element of Quine's account. Indeed, in Davidson's opinion—which is not easily discounted in this regard—"Quine emphasizes [the principle of charity] only in connection with the identification of the . . . sentential connectives" (Davidson 2001e, 136n16). So there is little to be said in favor of pushing logic-charity to the margins of Quine's account, as does Hookway.

It is true that Quine apparently makes a distinction between two stages of translation. The first includes the rendering of the truth-functional connectives, whereas the second involves filling out the translation manual with analytical hypotheses which fulfill requirements such as that to preserve stimulus-analyticity

(1960, 68). It would be a mistake, however, to suppose that these two stages correspond to a distinction between an obligatory, constitutive element and a non-binding, purely pragmatic one in Quine's account of translation, respectively. There is simply no indication that Quine regards the preservation of stimulus-analyticity as merely an optional pragmatic addendum to his main account of translation. Quine's does allow that the translator can turn an occasional blind eye to the requirement of preserving stimulus-analyticity if this permits substantial simplification in one's analytical hypotheses (1960, 69); and this, to be sure, represents a concession to the pragmatic consideration of simplicity. But it is not the requirement to preserve stimulus-analyticity (i.e., charity) which is the pragmatic element here; rather, charity is treated here as a generally *de rigeur* element that can *occasionally* be trumped by pragmatic considerations of simplicity. Again, there is a sense in which Quine's whole attitude to the intentional realms of mind and meaning is pragmatic and instrumental. Since his account of mind and meaning includes charity, perhaps this might be thought to warrant a view of charity as pragmatic. But Quine presents the non-factual, instrumental character of the intentional as an *implication*—on the basis of the indeterminacy of translation[51]—of what he takes to be the only plausible account of meaning, *not* as a presupposition of it. So Hookway's pragmatic interpretation draws no support from this quarter either. So much for the pragmatic interpretation of Quinean charity.[52]

[51] A word on the relation of charity and the indeterminacy of translation for Quine: Georges Rey observes that though Quine's argument for the indeterminacy of translation is often cited as a consideration in favor of normativism, the argument does not entail a principle of charity (Rey, 2001, 124n20). This is correct. Charity figures more as a crucial presupposition of the interpretative scheme that, in Quine's view, leaves scope for indeterminacy than as an implication of it.

[52] Issue may be taken as well with the *content* of the charity principle that Hookway wishes to ascribe to Quine. Hookway writes, "we are likely to prefer translations which maximize agreement between

It remains to consider the relative merits of the heuristic and constitutive interpretations. In the present context, we can take the heuristic interpretation as holding that Quine's principle of charity has an *a posteriori* status and that its content consists in the injunction not to attribute obvious falsehoods to agents unless there is strong evidence to suggest otherwise. The following three statements of Quine's represent the chief textual support for the heuristic reading:

> "The maxim of translation underlying all this is that assertions startlingly false on the face of them are likely to turn on hidden differences of language." "[O]ne's interlocutor's silliness, beyond a certain point, is less likely than bad translation . . . ." (1960, 59) "[T]he more absurd or exotic the beliefs imputed to a people, the more suspicious we are entitled to be of the translations . . . ." (1960, 69)

On their face, these statements do suggest that there is a presumption against attributing obvious falsehoods but that the presumption is defeasible given sufficient countervailing evidence. Moreover, as noted above (p. 30), Glock cites "Quine's naturalism" as counting against Quine's taking a constitutive approach to charity (2003, 33).

But these considerations must be weighed against others which support a constitutive reading. At least with respect to logic, I have suggested that Quine takes an uncompromisingly strong view of the content of charity. Quine presents the assumption that natives' beliefs and utterances respect logical norms not as defeasible in individual cases but as binding. The strength of Quinean logic-charity, then, conflicts with the heuristic reading.[53] Moreover, the *centrality* of logic-charity for

the aliens and ourselves: the best translation will be one that minimizes inexplicable error . . ." (1988, 136). Thus, Hookway attributes to Quine a holistic, maximizing principle of charity. But, as I have suggested above, Quine's principle does not appear holistic in general, nor maximizing in particular.

[53] In principle, charity's having a strong content is consistent with its having an *a posteriori* status. In that case, agents' universal respecting of logical principles would simply be an empirical discovery.

Quine, attested to by Davidson, favors the conclusion that charity is constitutive for Quine. Again, Quine's most explicit formulation of a charity principle—"Save the obvious" (1986, 82)—is free of any qualifying rider of the sort which the heuristic reading would append (e.g., "unless there is strong countervailing evidence"), despite the fact that Quine in this context stresses that the scope of charity extends beyond logic, taking in obvious truths in every domain, "every little bit of knowledge or discourse."

It cannot be denied that Quine makes statements that suggest a heuristic reading. But worth noting is that these statements mostly seem to function more in the way of backing or justification of his charity principle rather than as formulations of charity itself (cf. p.36 above). So there is a tension between Quine's application of charity (e.g., to logic) and his explicit formulation of it, on the one hand, and the claims he makes by way of justifying the principle, on the other. Whereas his charity principle (and application of it) itself seems quite strict, his efforts to justify it—such as they are—appeal to claims which are more lax. Perhaps faced with the prospect of providing no justification for charity at all, Quine preferred to make an appeal to "commonsense" (1960, 59), even though the facts commonsense certifies are not of sufficient strength to justify the sort of charity he actually seems to endorse and employ. At all events, in Quine's later statement of charity (1986, 82), it is the strict charity principle itself which is retained and not its problematic backing, of which Quine makes no further mention.

However, the only textual evidence that Quine might assign charity an *aposteriori* status is the very evidence for assigning it the content, associated with the heuristic reading by Stein and Glock, which sees attributions of logical coherence, etc., as defeasible assumptions.

On balance, then, the evidence inclines me to see Quine’s view of charity as constitutive. Thus, Quine can with some likelihood be classed as a genuine normativist alongside those like Davidson and Dennett whose interpretavistic views, though different from Quine’s in certain respects, bear the stamp of their predecessor in this as well as other respects.

Chapter Two: Normativism Within Davidson's Interpretationism

Introduction

In this chapter, I shall delineate and critically discuss more recent normativism, especially the influential version of it found in the work of Donald Davidson. In particular, my concern will be to begin to develop a characterization of the charity principles at which the arguments of later chapters will level. Although Davidson's version of charity will be the major focus of the chapter, comparison with Richard Grandy's *principle of humanity* will help bring out the form and function of Davidsonian charity, as well as certain liabilities to which it is subject. Finally, I shall examine the arguments—such as they are—with which Davidson (and others) have sought to defend charity as a constraint on agency.

Davidson's Interpretationism

Davidson enunciates charity principles in the context of expounding an interpretationist account of language and of the propositional attitudes. The *loci classici* for that account are the papers collected under the heading "Radical Interpretation" in Davidson (2001b), and that account is expanded in significant ways in Davidson (2004c).[1] A peculiarity of Davidson's interpretationist project is that its initial exclusive focus on language ultimately yields to a focus on intentionality more generally. The ascription of linguistic meaning to an agent's words comes to be seen as necessarily going hand-in-hand with the ascription to them of beliefs, desires, etc. But the later account, though more general, is also somewhat tentative and

[1] Davidson draws implications of his interpretavism and normativism for the possibility of psychological laws and physicalistic reduction in Davidson (1980c). See esp. 221-24.

programmatic. Davidson's more detailed (and better-known) formulations of his interpretationism actually occur in his treatment of language. So I shall devote considerable space to examining the role of charity in Davidson's account of language before I consider how, if at all, the proper appreciation of Davidsonian charity ought to be modified in the light of Davidson's more general interpretationist account.

Davidson's account of language is deeply indebted to two figures, Quine and Alfred Tarski.[2] In outline, Davidson's account, which he calls a theory of 'radical interpretation', strongly resembles Quine's theory of radical translation. In his account, Davidson is concerned chiefly with two questions: "What could we know that would enable us to" interpret someone's words on a particular occasion, and "How could we come to know it?" (2001e, 125). That is, Davidson is concerned with the questions in what sort of theory our knowledge of meaning might consist, and how we could acquire evidence for such a theory. Davidson plausibly insists that, although what we know must allow us to interpret a potentially infinite number of sentences, the underlying knowledge itself must be finite, given "that man is mortal" (2001f, 8-9). Moreover, the evidence for such a theory must not be semantic in nature (involving notions like meaning, synonymy, etc.) on pain of presupposing the very capacity of interpretation of which Davidson purports to give an account. Further, if one's aim is to explain the ability to grasp the meaning of utterances, then the account should not take the form of describing a translation method between languages, as it does for Quine. Such an account would allow one to "know which

[2] The extent of Davidson's debt to the former is reflected in his dedication to *Inquiries*: "*TO W. V. QUINE without whom not*" (2001b, v).

sentences of the subject language [the translating language] translate which sentences of the object language [the language translated] without knowing what any of the sentences of either language mean" (2001e, 129). As Davidson points out, one could use such a theory to interpret a language if the subject language happened to be one's own, but the interpretation of one's own language would necessarily escape the scope of such an account. Hence, in Davidson's view, it is not knowledge of a Quinean translation manual that underlies the ability to interpret a language.

Rather, Davidson submits, an interpretation theory might plausibly take the form of a Tarskian theory of truth, suitably modified so as to be able to handle the indexical elements ubiquitous in natural language. Such a truth theory respects the requirement of finitude and, Davidson writes, "entails, for every sentence *s* of the object language, a sentence of the form:

s is true (in the object language) if and only if *p*.

Instances of the form (which we shall call T-sentences) are obtained by replacing '*s*' by a canonical description of *s*, and '*p*' by a translation of *s*" (2001e, 130). A focus of much research has been whether the wealth of natural-language locutions permits treatment within the Tarskian framework.[3] But a more pertinent question which Davidson raises is whether (and how) one could confirm a truth theory for a natural language on the basis of the available evidence. For it is in this connection that charity enters into Davidson's account of language.

Davidson suggests, in effect, that standard hypothetico-deductive method can be used to test a truth theory. A proposed truth theory for a natural language can be

[3] Davidson (2001e, 132) gives the following partial list of potentially problematic locutions: "sentences that attribute attitudes, modalities, general causal statements, counterfactuals, attributive adjectives, quantifiers like 'most', and so on."

tested through the T-sentences which it entails: it is confirmed to the extent that it generates true T-sentences.[4] So the problem of testing such a truth theory reduces to that of determining the truth-values of a sampling of its generated T-sentences. Davidson finds a hint as to the solution of this problem if an interpreter can be supposed able to recognize when sentences are held true by members of a speech community.

Consider the (contextually-relative) T-sentence for the German sentence 'Es regnet' proposed by Davidson (2001e, 135):

> (T) 'Es regnet' is true-in-German when spoken by *x* at time *t* if and only if it is raining near *x* at *t*.

Davidson suggests that (T) could be confirmed (and, indirectly, any truth theory that yields it as a consequence) by garnering support for the following claim:

> (GE) (*x*)(*t*) (if *x* belongs to the German speech community then (*x* holds true 'Es regnet' at *t* if and only if it is raining near *x* at *t*))

If an ability to determine whether and in which circumstances individuals hold sentences true is conceded, it would indeed be a relatively straightforward matter to confirm—or falsify—(GE). Moreover, the truth of (GE) might seem to constitute strong evidence for (T).[5] But Davidson notes that because one can be wrong about

[4] In the light of his employment of Tarskian truth theory for purposes of interpretation, Davidson modifies Tarski's definition of a T-sentence. Tarski took the notion of translation for granted, and defined truth in terms of it. Davidson, by contrast, takes the notion of truth for granted and undertakes to define interpretation in terms of it. Accordingly, in the definition of a T-sentence, he understands *p*, not as a translation of *s*, but merely as a sentence true if and only if *s* is. Aside from allowing him to avoid a circular appeal to the very notion he is providing an account of, this permits one to test a truth theory for a language through recognition of correct T-sentences without presupposing a prior ability to interpret the language (as would be presupposed if *p* were required to *translate s*). Of course, there is a worry that with such an understanding of T-sentences, a truth theory cannot be expected to yield a genuine interpretation of a language. Davidson responds to this by expressing the hope that various constraints, formal and empirical, placed on a truth theory will suffice to render its T-sentences interpretative. I return to the subject of these constraints below.

[5] Its doing so, however, would seem to involve implicit appeal to some sort of charity principle, whether *a posteriori* or *a priori*. For one can readily imagine cases where, through some sort of

facts such as whether it is raining near one, one cannot "expect generalizations like (GE) to be more than generally true" (2001e, 136). So the bruited route to the confirmation of T-sentences does not quite pan out.

Davidson's Appeal to Charity

Precisely at this point, Davidson's discussion of the confirmation of a truth theory for a natural language takes a significant turn. Davidson abandons talk of confirming such a theory by means of confirming some sampling of its entailed T-sentences. Rather, he suggests a method wherein one aims for a "best fit." One chooses that interpretation of speakers' language that "maximizes agreement, in the sense of making" them "right, so far as we can tell, as often as possible" (2001e, 136). Davidson's idea is that an interpretation is preferable to the extent that the sentences speakers are seen to hold true actually turn out true (in the view of interpreters), when judged in accordance with the truth conditions (expressed in the entailed T-sentences) assigned to sentences by the interpretation. Thus, Davidson clearly subscribes to a kind of maximizing (or "optimizing") charity principle.[6] He maintains, however, that since there are an infinite number of sentences to consider, the maximization involved cannot be taken literally.[7] Moreover, Davidson introduces an important qualification on the principle. He writes, "it makes sense to accept intelligible error and to make allowance for the relative likelihood of various kinds of mistake" (2001e, 136). Thus,

perceptual inversion, a community holds true a sentence in circumstances precisely contradictory to its truth-condition. Again, one can easily imagine a community's holding true 'That is gold' co-varying with the presence of either gold or pyrite. But 'gold' might still refer to gold for all that, as Kripke teaches us. If sentences like (GE) are to serve as evidence for sentences like (T), such cases need to be rendered exceptional; and that would seem to require appeal to some sort of charity principle (more on charity presently, of course).

[6] See p. 45 above for a discussion of the general character of such maximization principles.

[7] Perhaps Davidson would have found the following understanding of the sort of maximization involved acceptable: That interpretation is to be preferred which renders true the greatest proportion of a large, representative finite sample(s) of sentences identified as held true.

it appears, some sentences incorrectly held true are not to be reckoned into the calculus of the optimal interpretation at all, and others are to be discounted in some degree according to what sort of mistake they represent.[8] At all events, charity clearly plays an important methodological role in Davidson's account of linguistic interpretation.

Davidson makes some tentative, broadly Quinean proposals concerning how one might plausibly *discover* a truth theory that interprets a speech community's language, though he does note some divergences between his and Quine's account. Like Quine, Davidson assigns a key role in the discovery of an interpretation to what Quine had called 'occasion sentences', namely, sentences assent to which is contextually-relative. In contrast to Quine, however, Davidson dispenses with the notion of stimulus meaning and takes assent to sentences' (general) co-varying with "objective features of the world" as a clue to their translation (2001e, 136).[9] Moreover, Davidson holds that, whereas Quine stresses charity "only in connection with the identification of the (pure) sentential connectives," he applies it "across the board" (2001e, 136). Davidson's point, apparently, is that the maximization principle he enunciates places a constraint on interpretations *in toto*, not just on the translation of sentential connectives. Indeed, the scope of Davidson's maximizing charity is broader than the logic-charity imposed by Quine's view of the translation of sentential connectives (cf. p. 34 above on logic-charity). It is concerned not only with

[8] Whether true normativists such as Davidson can comfortably allow for such qualifications to charity principles is an issue to which I return below.

[9] Again, appeal to charity seems implicit in Davidson's procedure here. That assent to 'Gavagai' co-varies with the presence of a rabbit can be taken to support the translation of 'gavagai' as 'Lo, a rabbit' only if hypotheses such as those mentioned above (p. 56, n. 5) are ruled out, which appears to require appeal to some sort of charity principle. However, the implicit role that charity serves here, in the context of discovery, is ultimately less significant than the methodological import which Davidson explicitly assigns it. So I do not dwell on it further.

the logical *coherence* of beliefs but also with the general *correspondence* of beliefs with the facts.[10]

Davidson makes some observations in defense of such a charity principle that shed considerable light on it. He writes,

> What justifies the procedure [of optimizing agreement with an interpreter] is the fact that disagreement and agreement alike are intelligible only against a background of massive agreement . . . . The methodological advice to interpret in a way that optimizes agreement should not be conceived as resting on a charitable assumption about human intelligence that might turn out to be false. If we cannot find a way to interpret the utterances and other behavior of a creature as revealing a set of beliefs largely consistent and true by our own standards, we have no reason to count that creature as rational, as having beliefs, or as saying anything. (2001e, 137)

This passage is of interest for several reasons. First, it leaves no doubt that Davidson regards charity as an *a priori* constraint on intentionality. Hence, whereas there was room for debate with respect to Quine, Davidson can unqualifiedly be accounted a normativist. Moreover, the passage reveals Davidson's adherence to a distinct Threshold Principle of charity (cf. p. 46 above). That is, intentionality, for Davidson, requires a certain—indeed, rather high—degree of truth and rational coherence among ones beliefs.[11]

In fact, in the passage quoted, Davidson—oddly—seems to appeal to such a Threshold Principle to justify his Maximization Principle. But the former principle is clearly unsuited to support the latter. For that a correct interpretation of an agent(s)

[10] In two respects, however, the contrast Davidson draws with Quine in point of the scope of the application of charity is misleading. First, as I have indicated in Ch. 1, though Quine's most conspicuous application of charity is in connection with the connectives, in fact, Quinean charity takes in far more than that. Second, though it is true that Davidson's maximizing charity constrains interpretations *in toto*, it should be borne in mind that he exempts "intelligible error" from the scope of charity.

[11] Presumably, however, Davidson intends the qualifications which he expresses concerning his Maximization Principle to apply to his Threshold Principle as well. That is, "intelligible error," etc., will be discounted in determining whether the threshold is achieved.

must render many (or most) of their beliefs true in no way entails that it must maximize truth relative to other candidate interpretations. In fact, Davidson must intend his employment of charity in the confirmation of an interpretative truth theory for a language to include his Threshold Principle as well as his Maximization Principle. An interpretation will be confirmed to the extent that it *both* renders a suitable proportion of the sentences which speakers' hold true, true in fact *and* surpasses other candidate interpretations in this regard.[12]

Indeed, Davidson's methodological proposal for confirming a truth theory would not be sound without the presence of the Threshold Principle. For if a theory is to obtain any significant hypothetico-deductive support, it must—together with auxiliary charity principles and observed facts about which sentences community-members assent to—yield entailments which are mostly true. Maximization by itself cannot guarantee this. At best, it can function methodologically to select among theories which *do* yield such entailments. So, despite initial appearances, the Threshold Principle is an indispensable element of Davidson's methodological proposals.[13]

The Significance of Charity in Davidson's Account of Meaning

So charity plays an important methodological role in Davidson's account of meaning. But more needs to be said about the general character of that account and the role of charity within it. One might get the impression initially that the import of

[12] This still leaves the Maximization Principle unsupported by Davidson's explicit statements, but at least Davidson "justifies the procedure" to the extent of appealing to Threshold charity as an *a priori* constraint on intentionality. Perhaps he would justify the maximizing element similarly.

[13] Might one hold that Davidson is advocating a distinctive, non-standard *hermeneutical* methodology? This seems unlikely since he undertakes to defend only the charity principle on which his proposal relies, not the methodological soundness of such reliance. Moreover, his initial consideration of whether a theory might be confirmed through establishing its T-sentences suggests acceptance of hypothetico-deductive canons.

charity in Davidson’s account of language is purely epistemological, that it enters in only to explain how one can confirm a translation of a language. But its significance is greater than that. For as Davidson points out, it is not enough that a truth theory yield correct truth-conditions for sentences to count as a theory of meaning for that language.[14] A sentence like “‘Snow is white’ is true iff grass is green” would count as expressing a correct truth-condition for ‘Snow is white’, but could hardly be regarded as giving the meaning of that sentence. Some means of guaranteeing that the right-hand sides of T-sentences genuinely translate the sentences mentioned on the left is required. Davidson speculates that a theory confirmed along the methodological lines he has sketched can be regarded as yielding T-sentences that are genuinely interpretative. He writes, “we have supplied an alternative criterion: this criterion is that the totality of T-sentences should . . . optimally fit evidence about sentences held true by native speakers . . . . If that constraint is adequate, each T-sentence will in fact yield an acceptable interpretation” (2001e, 139).

This statement is of the utmost importance in understanding the nature of Davidson’s account of meaning and the role of charity within it. For it makes clear that Davidson defines interpretation in terms of the evidence which would serve to confirm a theory of interpretation for a language. It is precisely this constitutive role that he assigns to evidence that renders his account of language interpretationist (cf. p. 14 above). Moreover, for Davidson it is accordance with his charity principle(s) that represents the key evidential element in his account of interpretation. A theory’s passing the test of charity is what is supposed to ensure its interpretativeness. Hence,

[14] Davidson understands the truth-conditions of a sentence very weakly, as captured by any sentence with the same truth-value as the original sentence.

charity plays a methodological role for Davidson, but for that very reason, given the nature of his account, it is constitutive of meaning as well.

Moreover, because of the role of charity within Davidson's account, Davidson cannot be taken to be giving a reductive analysis of meaning in terms of concepts "better understood . . . or more basic epistemologically or ontologically" (2001e, 137). "Concepts like those of meaning," Davidson writes, are ". . . not reducible to physical, neurological, or even behaviouristic concepts" (2001a, 154). Davidson's point is that an account of intentional notions like meaning cannot be given in wholly non-intentional terms. One, so to speak, is caught in an intentional circle.

At the end of the process of interpreting a language, Davidson holds that it is likely some indeterminacy will remain. But since he thinks that the constraints his account places on acceptable interpretations are more stringent than analogous Quinean ones, he thinks that the sphere of indeterminacy will be correspondingly smaller. In fact, he maintains that "the range of acceptable theories of truth can be reduced to the point where all acceptable theories will yield T-sentences that we can treat as giving correct interpretations . . ." (2001a, 152). Ultimately, he thinks there is an arbitrary but innocuous element of choice among schemes of interpretation, analogous to the arbitrary choice among differently calibrated scales of measurement. "Indeterminacy of this kind," he writes, "cannot be of genuine concern" (2001a, 154). Unlike Quine, then, by no means does Davidson wish to use indeterminacy to argue the ultimate illegitimacy of meaning or intentional discourse generally.

The Mutual Dependence of Meaning and Thought

Davidson, even in setting out the above account of the radical interpretation of language, is quite explicit that the account is provisional. For he holds that the interpretation of language and the attribution of propositional attitudes must go hand-in-hand: "... interpreting an agent's intentions, his beliefs and his words are parts of a single project, no part of which can be assumed to be complete before the rest is" (2001e, 127). Thus, in giving an account of the radical interpretation of meaning, one cannot take the attribution of attitudes for granted. An account of the interpretation of language cannot rely on evidence consisting of speakers' complex communicative intentions.[15] For Davidson maintains that it is impossible to establish the presence of such attitudes independently of someone's verbally communicating them. So availing oneself of this source of evidence would—illicitly—presuppose the ability to radically interpret itself. The problem is that "the attribution of thought depends on the interpretation of speech" (2001g, 163).[16] By the same token, it seems the interpretation of language cannot escape dependence on the detailed attribution of thought. Whereas in (2001e) Davidson gives the impression that the identification of one's holding a sentence true is unproblematic, his considered view is quite different: "there is no chance of telling when a sentence is held true without being able to attribute desires and being able to describe actions as having complex intentions" (2001g, 162).[17] In view of the mutual dependence of meaning and thought, then,

[15] Relatedly, Davidson also rejects accounts like those of Wittgenstein and Grice which attempt to cash linguistic meaning in terms of such intentions. Cf. (2001e, 127) and (2001a, 143-44).

[16] Davidson (2001g) argues for this claim at length.

[17] Additionally, Davidson comes to hold that an account of linguistic interpretation must take into account that the attitude of holding a sentence true, as a form of belief, admits of degrees. The necessity of identifying the degrees to which sentences are held true gives additional force to the claim that appeal to propositional attitudes is inescapable.

Davidson proposes a unified account of them. My discussion of that—rather technical—account is necessarily simplified.

Davidson presents his account in outline in (2004e). His aim is to sketch a theory that allows for the simultaneous interpretation of a speaker's language and attribution of beliefs and desires to them. His account draws heavily on decision-theoretic ideas. Frank Ramsey's Bayesian decision theory serves as a model of the sort of theory at which he is aiming. But in order to encompass meaning as well as belief and desire, he supplements Ramsey's theory with ideas deriving from Richard Jeffrey's version of Bayesian decision theory.

Davidson, following Ramsey, takes for granted that choices among courses of action (or preferences among states of affairs) are generally determined by the subjective probabilities and valuations agents assign to possible outcomes of alternative courses of action (or states of affairs) according to the principle that agents maximize expected utility. Thus, on the basis of someone's choices, if one were in possession of their valuations of various outcomes, one could compute their degrees of belief in them; similarly, if one were in possession of their degrees of belief, one could compute the valuations. Ramsey, however, proposed a clever way for computing *both* from their choices alone.[18] Thus, Ramsey provides a method for attributing beliefs and desires to an agent on the evidential basis of their "preferences between alternatives, some of them wagers" (2001a, 146)

[18] Davidson describes the technique so: "Ramsey solved this problem by showing how to find a proposition deemed as probable as its negation on the basis of simple choices only. This single proposition can be used to construct an endless series of wagers choices among which yield a measure of value for all possible options and eventualities. It is then routine to fix the degrees of belief in all propositions" (2004e, 153).

Davidson is very explicit with respective to the normative underpinnings of Ramsey's theory. It assumes a "reasonable pattern of preferences between courses of actions" (2001g, 160) and a rational coherence of one's values "in combination with [one's] beliefs" (2004e, 153). Davidson refers to the "conditions postulated by the theory" as "idealized" (2001g, 160). Evidently, the Ramseyan account of attribution relies on an extremely strong charity assumption.

But despite what Davidson regards as the merits of the Ramseyan account, Davidson notes that it is subject to the criticism that it would need to be supplemented by a theory of the interpretation of language. For, Davidson writes, "To learn the preferences of an agent, particularly among complex gambles, it is obviously necessary to describe the options in words. But how can the experimenter know what those words mean to the subject?" (2001a, 147). That seems to require a theory of interpretation on the part of the experimenter. But since the interpretation of language in turn seems to require detailed knowledge of propositional attitudes (cf. p. 63 above), we would be caught in a circle. As a way out of this circle, Davidson proposes a theory, incorporating elements of the Ramseyan account, that interprets language as well as attributing attitudes.

Moreover, Davidson abandons the idea, which in essence he had inherited from Quine, that agents' attitudes of holding sentences true suffice as an evidential basis for a theory of meaning. For various reasons, he thinks that knowledge of the degrees of agents' belief in sentences is required, but such knowledge is not readily gleaned by an interpreter. Davidson points out, however, that, on the one hand, decision theory seems to require a theory of meaning and, on the other hand, the theory of meaning seems to require a theory of degree of belief such as Bayesian decision theory can provide, suggests that the two are "evidently made for each other" (2004e, 158). But, again, to avoid circularity, a unified account of belief, desire, and meaning needs to be provided.

In that account, Davidson takes the attitude of an agent preferring one sentence true rather than another as basic. As he points out, such attitudes on the part of an agent can plausibly be seen as "a function of what the agent takes the sentences to mean, the value he sets on various possible or actual states of the world, and the probability he attaches to those states contingent on the truth of the relevant sentences. So it is not absurd to think that all three attitudes of the agent can be derived from sentences preferred true" (2004e, 158-59).

In outline, the unified account will have the following structure[19]: Degrees of belief in sentences, as well as comparative strength of desire that sentences be true, will be attributed on the basis of *preferences* that sentences be true. Meaning, in turn, will attributed on the basis of "knowledge about the degrees to which sentences are held true" (2004e, 159). Once the meanings of sentences are determined, of course, propositional belief and desire fall out directly.

Davidson emphasizes the crucial role of charity in his unified account of interpretation and meaning. He writes, "What makes the task practicable at all is the structure the normative character of thought, desire, speech and action imposes on correct attributions of attitudes to others, and hence on interpretations of their speech and explanations of their actions" (2004e, 166). The role of charity in this task sheds

[19] For Davidson's account of how such attribution might proceed, developed from Richard Jeffrey's version of Bayesian decision theory, see (2004e).

light on additional issues in the interpretation of Davidsonian charity to which I now turn.

Issues in Interpreting Davidsonian Charity

Among the issues with respect to the interpretation of Davidson's view of the propositional attitudes are the intended scope of his account of interpretation and of his principle of charity. When Davidson addresses how ascription of attitudes is to proceed in any detail, as in (2004e), given the nature of his account, his focus is at most on sorts of attitudes—beliefs, desires, possibly actions—which have a decision-theoretical bearing. His view of the ascription of other sorts of attitudes is unclear. Taking a cognitive theory of emotion as a model, perhaps one might think that other sorts of attitudes could be defined in terms of beliefs or desires. But recourse to this maneuver is pretty clearly ruled out for Davidson by the following quote:

> It is doubtful whether the various sorts of thoughts can be reduced to one, or even to a few: desire, knowledge, belief, fear, to name some important cases, are probably logically independent to the extent that none can be defined using the others . . . . (2001g, 156)

So the general ascription of attitudes other than belief and desire will not fall directly out of the account of (2004e) for Davidson. He *does,* however, write that "belief is central to all kinds of thoughts" (2001g, 156).[20] So, apparently, he views the ascription of other sorts of attitudes as parasitic in some way on the ascription of the decision-theoretical ones.

This has the result of making them at least indirectly subject to charity constraints. But beyond that, there is clear evidence that Davidson regards the principle of charity as a direct constitutive constraint on the attitudes generally. In

[20] Cf. (2004e, 152) where he refers to the decision-theoretic attitudes as "the central cognitive and conative attitudes" and as "basic intensional notions".

(1980c), he writes, “we make sense of particular beliefs only as they cohere with other beliefs, with preferences, with intentions, hopes, fears, expectations, and the rest” (1980c, 221).[21] So evidently Davidson holds that the sphere of charity encompasses all sorts of attitudes which are at all subject to rational and epistemic norms.[22]

A further issue of interpretation of Davidsonian charity concerns the *strength* of the constraints that it imposes on those attitudes that fall within its scope. In addressing this issue, some distinctions need to be carefully made, in the first instance, a distinction between (1) the strength of the rational (and epistemological) norms by which processes and products of thought are to be assessed with respect to their rational (or epistemological) propriety, and (2) the degree of adherence to such norms, of successful performance, which Davidson views as a necessary condition for the possession of agency. With respect to the former, there can be little doubt that the intended norms are of ideal strength.

A recent trend in the theory of rationality advocates that standards of rationality should be naturalized. In general, this is the view that “various empirical facts about humans and our environment must be taken into consideration in determining what the normative principles of reasoning are” (Stein 1996, 36). A central impetus behind this approach is the axiom that ‘ought’ implies ‘can’. If, as

[21] Cf. (2004c, 169-70) where Davidson notes that “The existence of reason explanations . . . is a built-in aspect of intentions, intentional actions, and many other attitudes and emotions . . . . An aura of rationality, of fitting into a rational pattern, is thus inseparable from these phenomena . . . .”

[22] Given the holistic nature of Davidsonian charity, there seems to be an at least *prima facie* tension between Davidson’s wish to subject attitudes quite broadly to charity and the need to treat of non-decision-theoretic attitudes in a second round of ascription parasitic on a prior round of ascription of basic ones. For this, in effect, isolates the basic ones from the latter—in violation of the holism of Davidson’s charity—in assessing the coherence of candidate sets of basic attitudes for purposes of ascription.

has seemed compelling to many, ethical norms are conditioned by possibility, then perhaps rational norms are as well, in which case rational norms would need to respect the limitations to which finite human agents are subject. But since such limitations could only be determined empirically, the knowledge of rational norms would be rendered, at least in part, *a posteriori*.

But such an approach, clearly, runs directly counter to the constitutive role which rational norms are meant to play in Davidson's account of the propositional attitudes. For Davidson is giving an account of the attribution of attitudes. However, the limitations which must figure in a naturalized theory of rationality would surely include fairly detailed facts about human beings' cognitive resources such as could only be gleaned through empirical knowledge of human psychology. But acquiring such knowledge presupposes the attribution of attitudes and, therefore, cannot explain it on pain of circularity. Naturalized standards, then, are plainly off-limits to Davidson.[23] So the norms which Davidson must appeal would need to be *a priori* and, therefore, inasmuch as they would not be qualified by empirical psychological limitations, in some sense ideal.

But whether Davidson requires ideal *performance* with respect to the relevant norms is a separate question. Christopher Cherniak (1986, 17-18) sees evidence in Davidson's "Psychology as Philosophy" (1980d, 237) that Davidson does require perfect consistency among one's propositional attitudes. For Davidson there seems to assume that the transitivity of preference cannot be coherently violated. Cherniak, further, cites Quineian charity as a source for Davidson's requirement of perfect

[23] There is the additional point that Davidson's central work on charity and radical interpretation mostly precedes the vogue of naturalized approaches to rationality.

consistency. Indeed, Cherniak correctly observes that Quine's view of translation in *Word and Object* presupposes perfect consistency for natives (at least with respect to their beliefs) (cf. p. 40 above).[24] But, Davidson's words in (1980d) notwithstanding, the bulk of evidence suggests that Davidsonian charity can brook lapses of consistency. Granted, Davidsonian charity includes a Maximization Principle, requiring that among competing interpretations that one is to be preferred which renders interpretees' attitudes truest and most coherent. But there is no suggestion that a correct interpretation must (or even can typically) achieve perfect truth and coherence. Rather, Davidsonian charity principles entail merely that an agent possess "a set of beliefs [and other attitudes] *largely* consistent and true by our standards" (2001e, 137) (cf. p. 59 above on Davidson's Threshold principle). So Davidsonian charity decidedly does not require ideal performance relative to rational norms and the norm of truth. Cherniak errs, then, it appears, in saddling Davidson with Quine's requirement of perfect conformity to such norms.[25] Given the evident implausibility of such a requirement, it is well that Davidsonian charity does not demand it.

[24] Moreover, with respect to what Davidson terms *Correspondence*, Quine assumes perfect performance as well: A native should never be attributed an obviously false belief. (Recall that Quine's norms are couched in terms of obvious truth, not truth *simpliciter*. So a native's entertaining a *non*-obviously false belief would not bear on one's performance relative to the respective norm.)

[25] Cherniak is clearly concerned to distinguish his own view, a moderate normativism (see below), from the views of paradigmatic normativists like Davidson and Quine. Even though the contrast Cherniak attempts to draw here with Davidson's view is errorneous, it is likely that when all is said and done, Davidson, in fact, impose more exacting charity constraints than does Cherniak: Davidson's Threshold Principle, though vague, seems to set the bar higher than does Cherniak's analogous principle.

Moreover, though Davidson does not require ideally rational performance, I shall suggest subsequently that Davidson *does* require an ideally rational *competence*.

Davidson's Charity versus Grandy's Humanity

Thus far, I have presented Davidson's Correspondence charity—as, indeed, Davidson himself often does, especially in his earlier papers touching on charity[26]—as a matter of the truth of a subject's beliefs, or, somewhat more accurately, their agreement with an interpreter's beliefs. But by the time Davidson writes his central papers on radical interpretation, possibly under the influence of Richard Grandy's "Reference, Meaning, and Belief" (Grandy 1973), Davidson appears to recognize the inadequacy of understanding Correspondence charity in these simple terms and to take the first steps towards a more careful formulation. Accordingly, in this section I discuss Grandy's critique of charity interpreted in terms of truth or agreement, and Grandy's attempt at a formulation of a more adequate alternative constraint, "the principle of humanity." Then I compare Davidson's more considered formulations of Correspondence charity to Grandy's principle. The excursus will afford considerable insight into Davidson's mature formulations of charity.

Grandy's critique is actually directed towards Quine's principle of charity, which Quine presents initially in his account of radical translation in *Word and Object* (1960). Grandy interprets Quine's principle of charity as dictating that in translation one should seek to maximize a subject's agreement with the translator with respect to obvious truths.[27] Grandy thinks the principle so formulated is correct but not general enough. Better, he notes, might be a more general principle which stresses that "the importance of agreement is proportional to obviousness" (1973, 441). But Quine ignores that obviousness is a matter of degree, as well as the fact

[26] Cf., e.g., (1980c).

[27] I, however, reject a maximizing reading of Quinean charity (see p. 46 above).

that obviousness is relative to a subject's "situation," where this includes factors such as "focus of attention, expectations, instrumentation" and "the past history of a speaker" (1973, 441, 443). Grandy observes that Quine possibly ignores such factors because they do not permit of ready behavioristic definition.

By contrast, Grandy proposes an alternative constraint on translation, namely, the principle of humanity, which takes these factors into account.[28] Interestingly, his account bears the stamp of the influence of Daniel Dennett's philosophy of mind.[29] Like Dennett, Grandy stresses the "pragmatic" purpose of translation and the ascription of attitudes. He sees the point of translation (and, it appears, interpretation) as allowing prediction and explanation of behavior. A successful translation can be used in determining an agent's beliefs and desires, which, with the help of "some model of the agent" can, in turn, be used to predict and explain its behavior (1973, 443).[30] But whereas Dennett's intentional stance essentially appeals to principles of rationality in order to link up actions to beliefs and desires, Grandy, with little in the way of argument,[31] rejects "mathematical decision theory" as a suitable model of the agent. Instead, he suggests that, in fact, we use ourselves as a model and determine others' actions by considering what we would do if we had the relevant beliefs and

[28] He plainly holds that his alternative constraint entails Quinean charity—re-interpreted to take account of degrees of obviousness and relativity to situation—as just described.

[29] Dennett's "Intentional Systems" (1971) appeared a couple of years before Grandy's article.

[30] It is tempting to see Grandy here as, like Dennett, espousing an instrumentalism. But it is not clear that insisting on the predictive value of translation and interpretation automatically renders one an instrumentalist. After all, Fodor, an arch-realist, would see the claim to reality of psychological posits as bound up with their role in laws and, therefore, equally closely bound up precisely with their potential use in prediction and explanation.

[31] Grandy allows that having elicited an agent's entire set of beliefs and desires, one might perhaps use decision theory to predict its behavior. But, he writes, "this is not what we do in practice" (1973, 443). His objection seems based on the implausibility of our actually relying on a whole system of attitudes as the basis of prediction.

desires.[32] That is, he emphasizes a role for simulation in our understanding of others. But, Grandy maintains, if the connections among attitudes (and the world) are insufficiently similar to ours, then we shall be unable to derive predictions in this way from their attitudes. So as a "pragmatic constraint on translation" Grandy proposes the principle of humanity, "the condition that the imputed pattern of relations among beliefs, desires, and the world be as similar to our own as possible" (1973, 443). Thus, without the right relations, no attitudes.[33]

But so formulated, the principle is vague. It leaves unspecified *what* are the relevant relations among one's beliefs, desires, and world that one must approximate in interpreting others. Grandy's idea *seems* to be that there are principles which capture how human attitudes must characteristically relate among themselves and to the world, and in choosing among translations one should choose that one that maximizes a subject's adherence to these principles. There is a question, however, whether these principles are to be thought of as normative ones. He remarks that epistemological principles play a large role in telling us whether a particular sentence can be reasonably attributed to a speaker as an interpretation of his utterance" (1973, 445). Moreover, the instance he cites of such a principle—what he terms 'a causal theory of belief'—seems to be normative. Though he does not attempt to formulate it precisely, the idea seems to be that speakers generally form beliefs about physical objects with which they have had some causal interaction, however indirectly. Of

[32] Or, more precisely, if we had *some* of their beliefs and desires, since Grandy thinks it is not plausible that prediction is based on an agent's full set of attitudes.

[33] Grandy clearly thinks we possess a capacity of simulation whose deliverances are faithful to these relations among attitudes and the world and which, therefore, we can use in generating accurate predictions of our own and others' behavior. But since the details and plausibility of this proposal are somewhat remote from the topic at hand, I shall hereafter focus only on the content of Grandy's principle of humanity, not his suggestion about how it is employed in detail.

course, as Grandy himself notes, this principle is closely related to Alvin Goldman's causal theory of knowledge (Goldman 1967, cited in Grandy 1973, 446n12). Essentially, Grandy's principle just restates Goldman's constraint on knowledge (justified true belief) as a constraint on belief generally. Ed Stein, however, interprets the principle of humanity as permitting among the relevant interpretative principles, principles of reasoning that diverge from normative principles (Stein 1996, 121), and certainly Grandy asserts nothing which commits him to the view that the principles constraining translation are all normative. At all events, I shall proceed on the assumption that Stein is correct in this regard.

Grandy argues the superiority of his principle of humanity over, in effect, simple Correspondence charity by considering cases where both intuition and the principle of humanity favor seeing a subject's utterance as reflecting a false belief but charity dictates presumptively viewing it as expressing a truth. Suppose, for example, that you are standing with Paul at a party and he utters the sentence 'John is a philosopher', having misheard the man standing nearby being called 'Ron' and referred to as a philatelist. As it happens, there is an individual in the garden named John (out of sight and earshot, and with whom Paul has not interacted either directly or indirectly) who happens to be a philosopher. Whereas charity would favor seeing the utterance 'John is a philosopher' as reflecting a true belief about the man in the garden, it is much more natural to see it as reflecting a false belief about Ron standing nearby. Indeed, if Grandy is correct that something like the "causal theory of belief" is among the principles constraining our attitudes, then the principle of humanity would seem to favor seeing Paul's utterance as reflecting a false belief about Ron, with whom he has interacted, rather than a true belief about John, with whom he has not. Moreover, Grandy maintains, “the principle of humanity . . . instructs us to prefer the interpretation that makes the utterance explainable. Since no reason could be given as to why Paul would have a belief about the philosopher in the garden, it is better to attribute to him an explicable falsehood than a mysterious truth” (1973, 445).

Davidson’s Response to Grandy’s Account

Whether or not Grandy’s argument is airtight, Davidson seems to have assimilated Grandy’s lesson that a correspondence principle couched in terms of maximization of truth or agreement will not suffice. Davidson writes, “The general policy . . . is to choose truth conditions that do as well as possible in making speakers hold sentences true when . . . those sentences are true. That is the general policy, to be modified in a host of obvious ways” (2001a, 153). The modifications which Davidson broaches in various places include the following: He writes that “it makes sense to accept intelligible error and to make allowance for the relative likelihood of various kinds of mistake” (2001e, 136); “Speakers can be allowed to differ more often and more radically with respect to some sentences than others, and there is no reason not to take into account the observed or inferred individual differences that may be thought to have caused anomalies . . .” (2001a, 152); “Disagreement about theoretical matters may be more tolerable than disagreement about what is more evident; disagreement about how things look or appear is less tolerable than disagreement about how they are” (2001g, 169).[34] What is striking about these

[34] Cf. from (2004e, 157): “agreement on what is openly and publicly observable is more to be favored than agreement on what is hidden, inferred, or ill observed . . . .”

pronouncements is the extent to which they suggest Grandy's principle of humanity.[35] Davidson's allowance of "intelligible error" is reminiscent of Grandy's injunction to "prefer the interpretation that makes the utterance explainable." Moreover, Grandy's claim (in response to Quine) that "the importance of agreement is proportional to obviousness" (see p. 72 above) corresponds very closely to Davidson's that "Disagreement about theoretical matters may be more tolerable than disagreement about what is more evident."[36]

However, whereas Grandy would defend such claims as these by an appeal to humanity, the requirement that "the imputed pattern of relations among beliefs, desires, and the world be as similar to our own as possible" (1973, 443), Davidson makes no (explict) appeal to humanity as an umbrella principle. Rather, his appeal is narrower, specifically, to epistemological considerations: "everything we know or believe about the way evidence supports belief can be put to work in deciding where the theory can best allow error, and what errors are least destructive of understanding. The methodology of interpretation is, in this respect, nothing but epistemology seen in the mirror of meaning" (2001g, 169).[37]

Davidson, then, unlike Grandy (see p. 63 above), *does* seem to limit the relevant "relations among beliefs, desires, and the world" to normative ones. His view seems to be that it is not truth or agreement of a speaker's beliefs with the

[35] As Ernie Lepore and Kirk Ludwig observe, "Davidson's conception of how the principle of charity is supposed to be applied is much more like what Grandy (1973) has called the 'principle of humanity' than his critics have supposed" (LePore and Ludwig 2005, 192n167).

[36] Inasmuch as Davidson, like Quine, appears to associate degree of obviousness with types of sentences, Davidson does not explicitly show quite the same recognition of the relativity of obviousness to a speaker's "situation" as does Grandy (see p. 72 above). But such a recognition would seem to underlie his statement that one should take into account "the observed or inferred individual differences that may be thought to have caused anomalies . . ." (Davidson 2001a, 152).

[37] For an interpretation of Davidson's Principle of Correspondence similar to the one I present here, see (Evnine 1991, 108-11).

interpreter's which the Principle of Correspondence enjoins one to maximize, but rather the extent to which the speaker's beliefs cohere among themselves, with the speaker's external circumstances and, presumably, with those of their cognitive mental states, such as sensations, which are not propositional attitudes.[38] Epistemic principles specify relations which must (not) obtain if beliefs are to possess epistemic warrant. Davidson's idea seems to be that one should minimize the extent to which the relations among these items violate relations dictated by epistemic principles. Grandy's 'causal theory of belief' can serve as an illustration as well as any putative principle. Thus, an interpretation which ascribes to a subject a belief about a physical object in the absence of (appropriate) physical interaction with it would detract from the coherence of the relations among their attitudes, circumstances, etc., for purposes of assessing the interpretation vis-à-vis other candidate interpretations. Generally, it is not falsehood (or disagreement) *per se*, then, which needs to be minimized, but rather lack of epistemic warrant.

But so understood, the distance between the Principle of Correspondence and the Principle of Coherence may seem to blur. After all, the former turns out to concern coherence in accordance with norms as much as the latter. To retain a distinction, one might try to appeal to the fact that in the case of Correspondence the relevant norms are epistemic whereas in the case of Coherence they are rational. But epistemology, at least insofar as it concerns relations of justification among beliefs (inference), pretty clearly overlaps with theoretical rationality. Ultimately, then, it

[38] Davidson's interpretationism, of course, is an account only of propositional attitudes. Unless he holds that mental states which are not propositional attitudes are somehow dependent on those which are, there seems no bar to Davidsonian charity (and, hence, his account of attribution of the attitudes) involving implicit reference to such mental states.

appears, Davidsonian charity seems to reduce to a single principle, which involves assessing the degree of coordination entailed by interpretations among beliefs, desires, cognitive mental states other than attitudes, and an agent's circumstances, in respect of both practical and theoretical rationality (as well as epistemology).

A Few More Interpretative Questions

Various interpretative questions, however, can be raised about the picture of Davidsonian charity I have just sketched. First, there is a question how well-defined charity constraints are on Davidson's accounting. Davidson himself writes, "It is uncertain to what extent these principles [i.e., "the rules for deciding where agreement most needs to be taken for granted"] can be made definite—it is the problem of rationalizing and codifying our epistemology" (2004e, 157). Indeed, given Davidsonian charity's dependence on principles of epistemic warrant (and of rationality), our current uncertainty as to the specific and even general character of such norms is bound to impart a certain indefiniteness to charity proposals. But even if our ignorance in this respect were abolished, there would still be a matter of spelling out the calculus to be used in assessing the extent to which candidate interpretations achieve epistemic and rational coherence. Such assessment, of course, is a pre-requisite for settling which interpretation is maximizing (Maximization Principle), and whether any even meet minimal requirements for agency (Threshold Principle).[39] In advance of precisifying the principle of charity, as Hans-Johann Glock notes, "charity becomes vacuous" (Glock 2003, 196).

[39] There is the additional important requirement of specifying at just what level the threshold is to be set. Davidson writes, "Making sense of the utterance and behaviour of others . . . requires us to find a great deal of reason and truth in them" (2001a, 153), but one naturally wants to know just *how much* is required.

Is a Requirement of Agreement Retained?

Question might be raised, moreover, whether Davidson, despite his revision of the Principle of Correspondence in the direction of epistemic coherence rather than truth or agreement, does not assign some role to the latter, nonetheless. In (2001g) where Davidson first *explicitly* characterizes Correspondence in epistemological terms, he still claims "what must be counted in favour of a method of interpretation is that it puts the interpreter in general agreement with the speaker" (2001g, 169). This might suggest that *both* epistemic coherence and agreement are to be weighed in applying Davidson's Principle of Correspondence.[40] However, in his "Introduction" to (2001b), Davidson expresses agreement with David Lewis that "Charity prompts the interpreter to maximize the intelligibility of the speaker, not sameness of belief" (2001c, xix). Moreover, by eliminating agreement from the purview of charity altogether, aside from avoiding—at least some—counterexamples of the sort that Grandy presents to charity, Davidson would seem to escape one of the perennial objections to interpretationist theories such as his.

As John Heil points out, an interpretationism like Davidson's appear subject to a regress problem. Such a theory seems to make the activity of interpretation part of the constitutive conditions of a mind's possessing propositional attitudes. Heil writes,

> The activity of interpretation itself, however, evidently involves interpreters' possessing propositional attitudes themselves. This points toward a regress . . . : my propositional attitudes depend on your interpreting me; your propositional attitudes depend on someone interpreting you; that someone's

[40] Evnine, despite appreciating the role of epistemic coherence in Davidsonian charity, implies that Davidson retains a place for agreement as well: "The point about interpreting people so that they come out believing truths, by the interpreter's lights (i.e. the Principle of Charity narrowly conceived), is simply one of the many principles which constitute Charity in its more developed sense" (1991, 110).

> propositional attitudes depend on some further someone; and so on. How could such a process get off the ground? (2004, 152).

But some care is required in considering the justice of this criticism as it touches Davidson's theory. It is true that Davidson's accounts of language and of the attitudes is shot through with references to 'interpretation' (he even labels his method for treating language, for example, 'radical interpretation', after all.) He couches these accounts in terms of the conditions which must be met for an interpretation of one's language or of one's attitudes to be acceptable. But 'interpretation' here is little more than a synonym of 'theory', I think, and introduces no more danger of a regress into Davidson's account than would employment of the latter word. Where a danger of a regress *does* arise is specifically with respect to the role of charity in Davidson's account. If the conditions for the acceptability of a theory of a subject's language or attitudes depends on meeting charity constraints, and ascertaining whether those constraints are met presupposes possession of a theory of one's own attitudes—as would be the case if Davidsonian Correspondence charity were a matter of agreement—then we would indeed be off on a regress. But if charity constraints make no reference to an interpreter's attitudes, as Davidson's revised principle of charity appears not to, then this risk of a regress is headed off.[41] Perhaps this consideration lends further credence to the reading of Davidsonian Correspondence as dispensing with a requirement of agreement among interpreted and interpreter altogether.

[41] Granted, the resulting account of attribution will involve implicit reference to the norms governing the various kinds of attitudes, and so there will be implicit reference to those kinds of mental states. But there will be no presupposition of *attribution*.

Could an Interpretationist Substitute Humanity for Charity?

As I have suggested, I believe that the evidence mostly suggests that Davidson's revision of Correspondence charity stops short of a full-fledged principle of humanity, in that he limits the relevant relations among attitudes, world, etc., to normative ones. It is illuminating, however, to consider whether it would, in principle, be open to an interpretationist like Davidson to adopt Grandy's humanity principle. First, it should be noted that the context in which Grandy proposes humanity as an improvement upon charity is different from that in which Davidson would have to employ it. For Grandy (1973), unlike Davidson, is concerned only with the presuppositions of radical translation, not with the methodology of the ascription of propositional attitudes as well. Thus, in spelling out the method of translation, he can take psychological ascription for granted in a way that Davidson cannot.

Moreover, it may be that employment of humanity presupposes psychological ascription and, therefore, cannot serve as a constraint upon it, on pain of circularity. This, in any case, is the view of Glock. As noted (see p. 73 above), application of humanity requires attentiveness to a speaker's "situation." For example, humanity dictates that one *not* translate a speaker as agreeing with one with respect to some particular belief if their situation is such as to afford no acceptable explanation for their having that belief. But Glock observes, "the principle of humanity faces an obstacle. How can we establish what the natives' position is, in advance of understanding any of their beliefs and desires?" (2003, 197). His point is that application of humanity presupposes ascription of the attitudes and, therefore, cannot without circularity serve in an interpretationist account of ascription such as Davidson's.

Some caution, however, is needed before accepting Glock's verdict in this regard. Glock seems to hold that application of humanity involves "crediting them [i.e., natives] with the beliefs and desires *we would have had*, if we had been in their position" (2003, 196). But, since their 'position' involves their attitudes, a temporal and logical priority of attribution is built into application of the principle. This, I think, can be questioned. As I have emphasized, what seems genuinely essential to humanity is "the condition that the imputed pattern of relations among beliefs, desires, and the world be as similar to our own as possible" (1973, 443). But to assess candidate interpretations with respect to how well they fulfill this condition, does not *obviously* involve any obvious prior determination of natives' attitudes.[42] Instead, it presupposes determining their situation *in the world* and seeing how well the attitudes postulated by particular interpretations cohere with one another and their situation so understood. If an interpretation fulfills the condition humanity sets, then the attitudes we would have (or actions we would undertake) "if we would have been in their position" (where this now *includes* their *other* attitudes) will allow us to make fairly reliable predictions about them. But this takes place *after* having made a humanity-based ascription of attitudes, not as a pre-condition of it.

Nonetheless, there are potential problems for a Davidsonian employment of a full-blooded principle of humanity in the not-too-distant offing. These concern the epistemological status of knowledge of the similarity of the relations among sets of attitudes and with the world to one's own. Grandy himself would, presumably,

[42] I consider below, however, whether it might in a less obvious fashion.

appeal to the deliverances of a faculty of simulation as the source of this knowledge (cf. Grandy 1973, 443). But aside from locating the source of this knowledge in a mysterious black box, this move re-instates the regress problem that Davidson's revision of Correspondence charity appeared to dissolve. For reading the output of one's simulator, so to speak, is itself an instance of attribution, *self*-attribution of beliefs about the coherence of attitude, etc. So attributing attitudes to others would presuppose their attribution in one's own case.[43] Because of simulation's mysteriousness and the risk of regress it gives rise to, an interpretationist of Davidson's sort should, I think, avoid appeal to it.

Instead of appealing to simulation, could a Davidsonian rely on an (at least partly) non-normative *theory* of the relations among "beliefs, desires, and world"? If this is conceived as an *empirical* theory, then the obstacle is readily apparent. The *a posteriori* confirmation of such a theory would presuppose having undertaken abundant attributions of attitudes. So, on pain of circularity, an account of attribution should not appeal to knowledge of such an empirical theory. Perhaps a Davidsonian *could* instead rely on a theory which was *a priori*. The idea is that it might be (partly) constitutive of the concepts of belief and desire that they (generally) be related in certain non-normative ways among themselves and with the world.[44] So, in principle, it appears that an interpretationist like Davidson could adopt Grandy's humanity principle. But, as noted, there is little textual evidence for ascribing such a view to

[43] Grandy himself can afford insouciance in this regard, since unlike Davidson, he is not concerned to give an account of psychological attribution. Davidson himself, one would think, needs to be more careful not to presuppose knowledge of one's own attitudes.

[44] A theory known as 'analytic' or '*a priori* functionalism' takes the view of belief and desire described in the text. Cf. p. 196 below.

Davidson (and, indeed, other prominent normativists). Accordingly, I shall not take it as the target for my arguments against normativism in subsequent chapters.[45]

An issue, though, is whether Davidson's revised charity principle, in which Correspondence is re-interpreted in terms of respecting epistemological norms, can really do the work that he asks of it. As mentioned (see p. 76 above), Davidson's revision of charity is prompted, at least in part, by a desire to allow "intelligible error" and do justice to Grandy's insight that "the importance of agreement is proportional to obviousness." Success in this would seem to depend, then, on whether errors are sufficiently explainable (and obviousness *cashable*) in normative-epistemological terms. Davidson's statement that "everything we know or believe about the way evidence supports belief can be put to work in deciding where the theory can best allow error" (cf. p. 76 above) suggests that he believes this to be the case. Whether this is plausible is an issue to which I shall return subsequently. In any case, that Davidson's revision of his Correspondence principle extends the ambit of charity beyond properly rational norms so as to include epistemological ones as well, will be critical in assessing the success of my arguments against Davidsonian charity in later chapters.

Arguments for Charity

As we have seen, all charity principles are not created equal. Some, such as Davidson's initial Correspondence principle, seem to have deficiencies, prompting proponents of charity to seek more acceptable alternative formulations. These

[45] Of course, a view based on humanity is, at best, misleadingly labeled a 'normativism'. For, in general, humanity does not assume that the typical relations among beliefs, desires, and the world possess a normative character, although specific proponents of humanity principles could hold that they do wholly or in part.

alternatives are perhaps more defensible to the extent that they avoid the defects of the simpler versions. But we have yet to consider what can be said by way of defense of charity principles (or, more properly, normativism) in general. Thus, I next consider what proponents of normativism have had to say in support of that view. On balance, normativists offer remarkably little in the way of argument for their view, and such arguments as they do advance, I point out, do not hold up under scrutiny.

Quine’s Arguments

The dearth of argument for normativism is most conspicuous in the case of Quine. In *Word and Object*, Quine’s ‘justification’ of his principle of charity seems limited to two sorts of things: (1) consideration of a few cases where intuitions might seem to support its application (for example, he observes that when someone answers ‘Yes and no’ to a question, we assume that the “the queried sentence is meant differently in the affirmation and the negation,” not that they are affirming and denying the same proposition [1960, 59]); and (2) an appeal to commonsense (“The common sense behind the maxim [i.e., the principle of charity] is that one’s interlocutor’s silliness, beyond a certain point, is less likely than bad translation” [1960, 59]). However, we can grant Quine that we are inclined, among other things, to translate so as not to attribute crass contradictions, indeed, that this inclination perhaps has a basis in “commonsense” along the lines he suggests. But that does not necessarily mean that that commonsense insight amounts to anything like an *a priori* constitutive constraint on belief. It could just as well represent recognition of the fact that people’s degree of “silliness,” as a matter of *a posteriori*, empirical fact, is generally fairly moderate. Indeed, the reference to likelihood in Quine’s statement of the "commonsense" underlying his maxim suggests a defeasibility that runs directly counter to the sort of strict, constitutive charity principle which, I have argued in Ch. 1, Quine ultimately champions. So Quine's discussion of charity in *Word and Object* does little to support normativism.[46]

Davidson's Arguments

Davidson's efforts to support charity, by contrast with Quine's, are a bit more substantial.

An Argument for Correspondence Charity

One line of argument proceeds as follows (2001g, 168-69). Davidson derives (1) Correspondence charity (for a given speaker "*most* beliefs are correct") from the premise (2) that a belief can have a particular subject matter ("object in, or aspect of, the world") only if one's beliefs about that subject matter are predominately true ("False beliefs tend to undermine the identification of the subject matter"). He endeavors to support this premise, in turn, by consideration of an example: (3) It is not clear to us that the ancients can be said to have believed that the earth is flat, since they lacked so many true beliefs about the earth (that "this earth of ours is part of the solar system, a system identified by the fact that it is a gaggle of large, cool, solid bodies circling around a very large, hot star . . .," etc.).[47] It follows, further, from (1)

[46] At (1986, 82), Quine's defense of charity consists in asking, "We have to base translation on some kind of evidence and what better?" Similarly, at (1960, 58), he asks "Not to be dogmatic about it, what criterion might one prefer?" This poses a legitimate challenge to opponents of normativism to articulate and defend an alternative conception of the propositional attitudes (and of translation) not based on *a priori* presuppositions of rationality, etc. The prospects of success in this, however, may not be nearly as bleak as Quine assumes.

[47] There is some complexity here that I am eliding. Davidson, it appears, suggests that the ancients' lack of many characteristic beliefs about the earth calls in question their having believed that the earth is flat because it calls into question whether their belief is about the earth (in the *de re* sense). He writes, "how clear are we that the ancients—some ancients—believed that the earth was flat? *This* earth?" (2001, 169). Of course, uncertainty of reference would yield uncertainty that they shared our concept of the earth, since sameness of concept entails sameness of reference. So it would not be clear

that (4) a theory of interpretation can be correct only if it is generally "the case that a sentence is true when a speaker holds it to be."

What are we to make of this argument? Well, Correspondence charity, (1), follows from (2), and, ignoring quibbles, (4) follows from (1). So the real question is whether Davidson's example adequately establishes the truth of the key premise, (2), and, if not, whether the truth of (2) is otherwise readily apparent. That the example does little to establish (2) becomes evident once the distinction between *uncertainty* and *indeterminacy* is borne in mind. One can concede to Davidson that the ancients' lack of many characteristic beliefs about the earth contributes to some uncertainty, psychological or epistemological, on our part as to whether the ancients can truly be said to have had a belief that the earth is flat. But (2), rather, plainly concerns the (degree of) determinacy with which individual's possess beliefs concerning a particular subject matter. For (3) to provide any real support for (2), it would have to be the case that uncertainty entailed indeterminacy. But that is far from the case. Our uncertainty as to the truth of Goldbach's conjecture, of course, in no way diminishes the determinacy of the truth of either that proposition or its contradiction.

Nor is the truth of (2) something that can be readily accepted as uncontroversial. For the trend in much theory of reference of recent decades has been to relax the dependence of linguistic—and, by extension, psychological—reference on an individual's possessing veridical information about a locution (or concept's) referent. Thus, though a descriptive theory of proper names like Searle's picks out

either that they believed that the earth is flat (in the *de dicto* sense). My feeling, however, is that there is relatively greater temptation to doubt that the ancients shared our concept *earth* than there is to doubt that they had a belief about the earth and its flatness (in the *de re* sense). But even the latter doubt is perhaps non-negligible. 'Earth' in the mouths of ancients might be compared to 'phlogiston' and lack a referent for similar reasons.

the referent of a proper name in terms of an individual's maximally *satisfying* descriptions predicated of that proper name, thereby tying reference to truth, a causal theory like Kripke's (at least for proper names and natural kind terms) appears to divorce reference from truth. Granted, if the sphere of application of such causal theories were relatively narrow, their truth, although qualifying Davidson's premise, would not seriously undermine it. But many proponents of such theories cast their net quite widely. So the truth of (2) is quite controversial, and, consequently, Davidson's use of it renders his argument for Correspondence charity unavailing.

It is worth noting, moreover, the limited scope of the argument's conclusion, even if it had succeeded in proving that conclusion. First, it concerns Correspondence charity alone, not Coherence charity. Granted, that truths predominate among one's beliefs entails a certain degree of coherence among one's beliefs, inasmuch as truths cannot contradict each other.[48] But it has no implications at all for practical rationality, which is supposed to be included in the domain of Davidson's Coherence charity. Moreover, with respect to theoretical rationality,

[48] It is possible, however, that Davidson would see its implications with respect to theoretical rationality as greater than this. Davidson's view that "a belief is identified by its location in a pattern of beliefs" and that this pattern "determines the subject matter of the belief" (2001g, 168) might suggest that he would hold that a belief's logical content is similarly dependent on the presence of an appropriate pattern of beliefs. So perhaps one cannot have a belief involving some logical operator like conjunction, a belief of the form *p & q* without having the beliefs *p* and *q* separately. That is, in general, he might be inclined to hold that having beliefs generally requires believing their deductive consequences. This would extend the implications of the conclusion of Davidson's argument for theoretical rationality beyond mere consistency. Perhaps Davidson might even take the relevant patterning to include not just what background beliefs exist but the inferential relations among beliefs as well. In that case, Davidson's view here might seem to entail that most of one's deductive inferences be valid as well (it is difficult to see how one could derive implications for non-deductive inferential relations). However, my discussion in the text is concerned only with the implications of the conclusion of Davidson's argument, not with those of ancillary views of his.

though it entails some statal rationality (consistency),[49] it entails nothing with respect to procedural rationality, such as the propriety of one's logical inferences.

Moreover, even with respect to Correspondence charity, Davidson's conclusion does not seem to deliver quite everything that Davidson would want. For it seems to concern only that part of Correspondence having to do with the truth of one's beliefs, not that having to do with the observance of epistemological norms governing warrant. As my reading of Davidson's revised Correspondence principle suggests, Davidson appears to intend that principle to require, not just that an agent's beliefs be mostly true, but that they have been mostly formed in ways that confer epistemological warrant on them. The former, however, does not entail the latter. However unlikely, it would seem logically possible for someone to have mostly true beliefs which lack warrant because formed in epistemologically dubious fashion.[50] Moreover, Davidson appears oblivious to the fact that the conclusion of his argument partakes only of the character of a Threshold Principle, asserting that the proportion of an agent's beliefs which are true cannot fall below a certain (high) level. In no way does it amount to a Maximization Principle, enjoining one to prefer that interpretation which renders the largest proportion of speakers' utterances true. But it is precisely Maximization charity which Davidson wishes to rely on for methodological purposes.[51] So his argument falls short of the mark in numerous ways.

[49] Even here, it does not encompass relations of non-deductive coherence among beliefs.

[50] Conversely, it appears logically possible that someone could be so singularly ill-situated that most of their beliefs, though warranted, are false. So the two sides of Davidson's revised Correspondence principle are logically independent.

[51] Davidson himself, just after presenting his argument, refers to Maximization as "the basic methodological precept" (2001g, 169). He appears not to notice the distance between the argument's conclusion and that imperative.

A Transcendental Argument

A second line of argument, for which some commentators find evidence in Davidson, appeals to charity, both Correspondence and Coherence, as a necessary condition or transcendental requirement for radical interpretation.[52] Lepore and Ludwig summarize the argument as follows:

> (1) Interpretation from the standpoint of the radical interpreter is possible.
>
> (2) If interpretation from the standpoint of the radical interpreter is possible, then the principle of charity is true.
>
> (3) Therefore, the principle of charity is true. (2005, 204)[53]

We have already seen (p. 58 above) that Davidson appeals to charity as a methodological device which is supposed to make it possible for the radical interpreter to confirm interpretations. If the commentators are right, his view is that charity is not only sufficient for this purpose (in the presence of certain other *a priori* and empirical constraints) but necessary, in that no other methodologically adequate expedient would be available to the radical interpreter. Whether recourse to charity is necessary (or even whether it is sufficient) for radical interpretation is something on which I shall venture no opinion.[54] But even if the second premise could be conceded, the first would itself be far from obvious. It is by no means clear that interpretation purely on the basis of such evidence as is available to the radical interpreter is possible. Barring extreme skepticism, one must grant that interpretation is a quotidian event. But ordinary interpretation is not restricted in its evidential

[52] Perhaps Davidson implicitly makes the argument in the following passage (2004e, 157): “Further interpretation requires the assumption of further agreement between speaker and interpreter. The assumption is certainly justified, the alternative being that the interpreter finds the speaker unintelligible.”

[53] (Glock 2003, 195) seems to find much the same argument in Davidson.

[54] An interesting question, however, is whether an appeal might be made to a principle of humanity instead of charity.

bases as is radical interpretation. As Ludwig (2004, 354) points out, “We can appeal to knowledge of features of our own psychological type and to the fact that in practice others whom we want to interpret are conspecifics embodied in the same way we are and in similar environments, to infer with some plausibility the sorts of things they are apt to be thinking, in order to constrain our interpretations.” So the fact of interpretation does not entail the fact of *radical* interpretation. Since the very possibility of radical interpretation is far from obvious, the soundness of Davidson’s argument for charity founded upon it is equally so.[55]

In a sense, it may not be altogether incumbent on Davidson to give anything like a traditional formal philosophical argument for his principle of charity. It might be enough if his account of the propositional attitudes, with its normativist underpinnings, were over the long run to prove the best explanation of the relevant phenomena by comparison with other initially plausible philosophical accounts of the attitudes. That is, perhaps the truth of Davidson’s interpretativism and its attendant normativism can be established by an inference to the best explanation.[56] However, on any account of abduction, the most central good-making property of a theory is its explanatoriness, roughly, the degree to which, on the assumption of its truth, it would lead one to expect the occurrence of phenomena in the relevant domain. In effect, my argument of later chapters can be viewed as suggesting that normativist theories like Davidson’s quite crassly violate the requirements of explanatoriness in that they entail the scientific impossibility of what are clearly possible phenomena. So,

[55] In fact, if radical interpretation has the implication that intentionality is indeterminate, as Davidson—following Quine—maintains, that itself should constitute some grounds to question the possibility of justified radical interpretation, given the *prima facie* determinacy of our thoughts.
[56] Cf. the discussion of this line of argument in (LePore and Ludwig 2005, 202-04).

although I am sympathetic to the employment of inference to the best explanation as an argumentative strategy with respect to issues in the philosophy of mind and elsewhere, I hold that this strategy ultimately redounds to the detriment of normativism, not in its favor.

An Argument from Dennett

Some commentators find a separate line of argument for normativism in Daniel Dennett's work. Dennett, clearly a proponent of normativism,[57] gives an instrumentalist account of the propositional attitudes. To possess attitudes is, on his view, to be an *intentional system*, "a system whose behavior can be (at least sometimes) explained and predicted by relying on ascriptions to the system of beliefs and desires (and hopes, fears, intentions, hunches, . . .)" (1971, 87). Thus, Dennett intimately connects possession of the attitudes to the pragmatic values of explanability and predictability. Moreover, Dennett takes for granted that explanations and predictions in terms of the attitudes will be rationality-based: "One predicts behavior in such a case by ascribing to the system *the possession of certain information* and by supposing it to be *directed by certain goals*, and then by working out the most reasonable or appropriate action on the basis of these ascriptions and suppositions" (1971, 90). It would seem to follow, then, that possession of the attitudes by a system for Dennett requires at least that degree of adherence to rational norms (Coherence) as will permit some explanation or prediction in accordance with

[57] Cf. esp. his statement that "a false belief system is a conceptual impossibility" (1971,101).

them. So the argument, if successful, establishes a Threshold Principle for Coherence.[58]

But just how high a threshold would it establish? In order to answer this question as well as to consider whether the argument succeeds, it will be helpful briefly to examine the version of normativism Christopher Cherniak expounds in his *Minimal Rationality* (1986). Cherniak's normativist outlook comes to the fore even in the first sentence of that work: "The most basic law of psychology is a rationality constraint on an agent's beliefs, desires, and actions: No rationality, no agent." (1986, 3). He makes it clear that he is "concerned with rationality conditions on belief sets and on the believer's deductive abilities," which he regards as "necessary conditions on agenthood" (1986, 5).[59] Moreover, it is clear that he accepts the necessity of charity constraints on much the same basis as does Dennett. He describes a theory of belief (the "*assent theory of belief*") which places no rationality constraints on agents: "*A* believes all and only those statements that *A* would affirm" (1986, 6). Cherniak

[58] A distinct line of thought in Dennett, which superficially might seem to be directed towards establishing normativism, on closer inspection is seen to level at a different, although related, conclusion. Dennett appears to hold that in environments in which natural selection operates, organisms will be explainable from the intentional stance, and so whatever amount of charity such explainability entails will apply to them (cf. 1971, 92-93). Thus, human beings' conformity to such charity is given by the *a posteriori* fact that they have been formed as a species in an environment in which natural selection operates. This conclusion falls short of normativism in two respects. At best it establishes that charity is a constraint on human beings (and other organisms) which have developed in this environment, not that charity is a constraint on possession of the attitudes. Moreover, Dennett's line of thought establishes its conclusion only *a posteriori*, whereas normativism, as I have defined the notion, regards charity as an *a priori* constraint.

[59] Cherniak's reference to charity, in effect, as a "law of psychology" might suggest that he regards charity as undergirded by an empirical generalization about agents' rationality. In that case, his view would not properly count as normativist. However, Cherniak emphasizes that although "rationality conditions perhaps are not usefully regarded as 'definitional', they must be distinguished from empirical generalizations about human psychology." He writes, further, that they "have a centrality in a cognitive theory, such that they could not be rejected on the basis of just some supposedly contrary 'data' (1986, 27). Thus, Cherniak appears to accept something like Quine's rejection of the analytic-synthetic distinction and accord the principle of charity the status that Quine and other adherents of charity influenced by him seem obliged to assign it, namely, that of a principle possessing "centrality" in the 'web of belief'. Accordingly, like Quine, Davidson, et. al., Cherniak can be comfortably reckoned a normativist.

rejects this theory essentially because of its lack of explanatoriness: “A cognitive theory with no rationality restrictions is without predictive content” as to “a believer’s behavior” (1986, 6). Such a theory, he maintains, would deprive one of an adequate evidential basis for making attributions of the propositional attitudes; and it renders mysterious our ability to make successful predictions on the basis of such attributions.

Cherniak, however, is quite explicit that such requirements of predictiveness only ground a moderate normativism. He holds that for both “everyday psychological explanations of behavior” and for “cognitive theory,” what is required is what he calls “minimal, as distinguished from ideal, rationality” (1986, 3). Whereas an ideal rationality, broadly, requires an agent to undertake all actions and inferences appropriate to their belief-desire set (and to refrain from all that are inappropriate to it), minimal rationality demands only that they undertake (refrain from) some.[60] Cherniak supports this weaker constraint by means of “exhaustion of a trichotomy” (1986, 8-9): (1) Because human beings are finite beings with finite cognitive and temporal resources, they could not satisfy a requirement of ideal rationality. Imposition of such a constraint would have the unacceptable implication of making cognitive theory inapplicable to human beings. (2) As argued above, if agents were subject to no charity constraint at all, then this would unacceptably undermine the predictive power of belief- and desire-attribution. Accordingly, (3), agents must be subject to a requirement of minimal rationality.

So Cherniak, sharing Dennett’s insistence on the necessity of charity constraints to assure the explanatoriness of attribution of the attitudes, argues to a

[60] Similarly, ideal rationality requires weeding out all inconsistencies from one’s belief-set, whereas minimal rationality merely requires weeding out some.

Threshold principle which sets the bar for possession of the attitudes below the level of perfect rationality. He holds that explanatoriness can be assured with a requirement of mere minimal rationality. But it is worth noting that as Cherniak specifies the contrast between ideal and minimal rationality, Davidsonian charity actually might seem to constitute a species of minimal rationality. For Davidson does not require perfect rationality of an agent, rather, merely a large degree of rationality.[61] Though Davidson, unlike Cherniak, might not have recognized the bar to perfect rationality for finite creatures which Cherniak emphasizes, the threshold Davidson sets for agency seems to fall in the range of minimal rationality. Moreover, as Cherniak himself acknowledges (cf.1986, 18-20), his minimal-rationality constraint is vague inasmuch as he does not specify a definite cut-off point for agency in the range between zero and perfect rationality. If minimal rationality suffices to guarantee explanatoriness and predictiveness, as Cherniak holds it does, then one would think that a reasonably high degree of adherence to rational norms would be required. Otherwise, predictions based on such norms will usually fail, and surely it is a requirement of genuine explanation that an *explanans* should predict its *explanandum* with a high degree of probability. Indeed, Cherniak himself evidently holds that the determinateness of one's qualifying as an agent is in proportion to one's degree of rationality. So, on closer inspection, Cherniak's view of charity seems to merge with Davidson's.[62]

[61] Cherniak exhibits some tendency to view Davidson as, in fact, requiring perfect rationality (1986, 17-18), but, as noted above (p. 70), I think the bulk of the textual evidence counts against such a reading of Davidsonian charity.

[62] Cherniak, however, appears to understand the sort of quantification involved in assessing whether a potential agent meets minimal requirements differently than does Davidson in at least one respect. Whereas Davidson would seem to make the assessment on the basis of something like the proportion

But how successful is Cherniak's (and, thus, Dennett's) argument for his view of charity? Overall, I think not very. I believe one can grant the first premise of his trichotomy argument, that agents are not subject to a requirement of perfect rationality for the reason he cites: that such a requirement would debar finite creatures like human beings from counting as agents. Moreover, if one grants his second premise, that agents must be subject to *some* charity constraint, his conclusion that they are subject to a minimal rationality constraint would be unavoidable. So the issue is whether the second premise must be conceded.

As noted, Cherniak grounds his second premise, the requirement of some rationality constraint, on the necessity that attribution of the attitudes make possible prediction and explanation. He supports this necessity, in turn, on the need to have an adequate evidential basis for attribution, and the need to do justice to our apparent success in making predictions on the basis of attributions. Let us grant for the moment that it is, in fact, crucial that attribution allow prediction and explanation. Nonetheless, it is not clear that this entails anything like a charity constraint as this is ordinarily understood. For, in the first place, it seems enough to guarantee the predictive and explanatory power of attribution if most agents should be largely rational. Their behavior could be predicted by applying rational norms to their set of beliefs and desires, even if such prediction failed in the case of a minority of oddballs

of *token* rational inferences (actions, etc.), Cherniak seems to wish to do so on the basis of the proportion of *types* of rational inferences which a potential agent regularly makes.

who exhibited a significant degree of irrationality.[63] So predictiveness, it seems, can be had without a charity principle binding on all agents.

Moreover, it is not at all clear that the basis of intentional prediction and explanation of agents need be (wholly) normative in character. For, as we saw (p. 72 above), Grandy shares Dennett's (and Cherniak's) emphasis on prediction and explanation of behavior as the pragmatic point of attribution. But Grandy, unlike Dennett, appeals to humanity rather than charity as rendering such prediction and explanation possible. As I have pointed out, humanity requires only that the relations among attitudes (and the world) be "as similar to our own as possible" (Grandy 1973, 443). In general, there is no assumption that these relations are normative, and, thus, none that the intentional prediction and explanation their obtaining make possible are normative in character. Cherniak himself evidently includes non-normative psychological principles in addition to normative ones as constraining attribution. He writes, "the minimal rationality conditions in everyday practice are embedded in a broad range of other cognitive psychological theories that fill in where the minimal agent's behavior will depart from ideal rationality" (1986, 55). He regards these non-normative theories as enshrined in commonsense psychology, enjoying a status as central as that enjoyed by normative principles. But the acknowledgement of such non-normative principles opens up the—at least theoretical—possibility that predictiveness could be afforded by such principles rather than by normative ones, a fact which calls into question the soundness of Cherniak's inference from the necessity of predictiveness to a principle of charity.

[63] Put another way: Once it is granted—as one must—that explanation and prediction need not be on the basis of strict, exceptionless laws, there opens up the possibility of individuals who are massively exceptional.

Further, Cherniak seems to assume that the principles, whether normative or non-normative, which render the attribution of attitudes predictive need to have a quasi-constitutive character.[64] But why cannot the predictive import of a theory of the attitudes instead be a matter of *a posteriori*, empirical generalizations whose confirmation the theory makes possible? Cherniak, apparently, makes a tacit assumption that predictive content needs to be, as it were, written into the theory of the attitudes itself, or else there will not be a sufficient basis for attribution. Indeed, if the possibility of attribution is undercut, the very possibility of confirming empirical generalizations in which the attitudes figure will be undercut. But it seems that the possibility of attribution of the attitudes can be allowed for without building predictive content *directly* into the theory of the attitudes. A Kripkean, scientific essentialist theory of the attitudes, such as an *a posteriori* functionalism, would—one would think—allow attribution to proceed on the basis of rough-and-ready reference-fixing descriptions of the attitudes (laying no claim to being quasi-constitutive),[65] and it would allow generalizations which permit reliable predictions on the basis of the attitudes to emerge only *a posteriori*. If such a scientific essentialism has seemed promising with respect to other scientific domains, then why not with respect to psychology as well? I conclude, then, that Cherniak's argument for normativism, just as the other typical arguments for normativism considered above, fails to establish its conclusion.

[64] This is apparent in his remark that "A cognitive theory [i.e., a philosophical theory of the attitudes] with no rationality restrictions is without predictive content; using it, we can have virtually no expectations regarding a believer's behavior" (1986, 6).

[65] Similarly, one can have considerable success in identifying specimens of gold and other minerals in the absence of possession of knowledge of (quasi-)constitutive constraints on the various sorts of minerals.

The upshot is that normativism—whether of Quine’s, Davidson’s, or Dennett’s sort—is left with little in the way of positive support. More needs to be said, however, by way of rounding out the picture of Davidsonian charity that I have drawn in the present chapter. It is to this task that I turn next.

Chapter Three: Normativism and Modules, *Prima Facie* Objections

Introduction

Although Davidson’s explicit discussions of charity advert to principles of the sort that I have treated hitherto—namely, Maximizing and Threshold Principles—in the present chapter I shall highlight charity principles of a different sort that Davidson appears to embrace, albeit somewhat less explicitly than he does the aforementioned principles. These principles, which I shall refer to as the Competence and Compartment Principles, respectively, appear to underlie Davidson’s account of familiar forms of irrationality like *akrasia* and self-deception in such papers as (2004c) and (2004a). That account sees a need to posit sub-divisions of the mind in order to allow for irrational mental processes and activities. Preparatory to developing my argument of Chapter Four, I shall find it helpful to discuss the relation of such Davidsonian compartments to the sorts of mental ‘modules’ that figure in much contemporary psychological theorizing. Distinguishing among various conceptions of modules, I shall argue that the properties of one important sort of module bear directly on the tenability of Davidson’s Competence and Compartment Principles. Last, I shall attempt to forestall certain objections that might be raised to my employment of the notion of modularity in arguing against normativism. In sum, my goal here is to clarify the relevance of what might be called the *mereology of the mind* to the issue of normativism.

Additional Charity Principles

There is good reason to believe that Davidson advocates charity principles besides those which he enunciates in his articles of the '70s in which he sets out his account of radical interpretation. These additional principles come to the fore in articles Davidson published in the '80s specifically concerning the topic of irrationality. Ed Stein claims—I think rightly—to discover in Davidson (2004b) a principle of charity concerning an agent's "reasoning competence" (Stein 1996, 116n14). Roughly, the principle makes a condition for their possessing propositional attitudes that for each rational norm, they embody a tendency to obey it.[1] But to gain a proper appreciation of the principle Stein has in mind some discussion of the notion of competence is required.

Competence and Performance

As Stein notes, the notion of a competence is closely associated with Chomskyan linguistics. Chomsky distinguishes between a speaker's linguistic *performance*, their actual linguistic behavior, and their linguistic *competence*, the (tacit) knowledge of linguistic rules—phonetic, grammatical, and semantic—which underlies and, along with other factors, serves to explain their linguistic behavior.[2] This distinction has the merit of accommodating the evident fact that one can have mastery of a language, knowledge of its rules, etc., and yet err in one's linguistic performances, such as speech production, comprehension, and one's intuitions about

[1] Stein speaks more narrowly of "principles of reasoning" because his focus is exclusively theoretical, but Davidson's is not: If (2004b) contains a principle such as the one Stein identifies, it concerns all forms of rationality. So it is appropriate to formulate it in terms of rational principles in general, as I have done here.

[2] A subtle question is whether the relevant knowledge should be thought of more on the model of 'knowledge how' or 'knowledge that', i.e., propositional knowledge. The word 'competence' itself is suggestive of the former. But Chomsky, at least in the first instance, may have had the latter more in mind.

grammaticality. Such performance errors are attributable to factors extraneous to one's linguistic competence. Stein distinguishes among factors that are "due to basic facts about the human condition" (such as "constraints on processing time and memory") and ones arising from one's idiosyncratic state (for example, "lack of attention" owing to "an inadequate amount of sleep, excessive drug use, or excitedness" (1996, 40).[3]

There is a complication that is glossed over by the simple identification of linguistic competence with tacit linguistic knowledge. Namely, as Stein points out, there are various interpretations of what a competence is. Aside from the interpretation of competence as a matter of knowledge, another viable interpretation sees it as the language-specific cognitive *mechanisms* which underlie one's linguistic performance (1996, 53-55). Whereas on the knowledge view competence is a matter of cognitive *states,* on the mechanism view it is more a matter of the cognitive *processes*, the transitions among psychological states which support language specifically. These transitions are naturally thought of as *ceteris paribus* regularities, and, correspondingly, on this view performance errors are naturally viewed as divergences from these regularities owing to interfering factors, whether physical or psychological.

What appears true of language, Stein points out, one might take to be true of reasoning as well. One might see individual normative principles of reasoning like *modus ponens* as embodied in a rational competence, which is only imperfectly

[3] Stein designates the latter "situational" factors and the former "psychological" ones. The latter designation is misleading since—as his examples themselves show—he clearly intends to include peculiarities of an individual's psychological state (like lack of attention) among situational factors. Stein's typology of the sources of performance error, then, appears substantially to cross-cut the classification of them as physical or psychological.

manifested in one's actual reasoning performance. Again, this view would admit of either a knowledge or a mechanism interpretation. On the former, one's reasoning competence would consist in one's knowledge of principles of reasoning like *modus ponens*, whereas on the latter, it would reside in regularities holding true of one corresponding to principles of reasoning, such as that if one believes *p and q*, as well as *p*, then, *ceteris paribus,* one infers *q*.[4] Again, on the latter view, divergences from the normative principles of reasoning would be seen as performance errors attributable to outside factors interfering with the mechanism's operation.

Availing himself of this notion of competence, then, Stein attributes to Davidson a principle of charity according to which "the principles of reasoning embodied in our reasoning competence" are "basically rational" (1996, 116). Let us call this charity principle Stein discovers in Davidson the *Competence Principle*.[5] Stein distinguishes strong and weak versions of this principle. On the strong version, agents "should *never* be interpreted as irrational" in the sense that "all divergences from the normative principles of reasoning should be classified as performance errors" (1996, 116). On the weak version, by contrast, agents should be seen as rational "*unless* there is strong empirical evidence to the contrary." However, the constitutive role charity plays in Davidson's account of attribution places the weak version out of bounds for his purposes.[6] Accordingly, I shall confine consideration to

[4] I suggest below that the mechanism interpretation more nearly captures Davidson's view.

[5] Stein—apparently correctly—further observes that this entails what he calls the "rationality thesis," that although "human beings can make errors in reasoning," these are mere external interferences with an ideal rational competence (1996, 3). This latter thesis, however, though relevant to Stein's purposes, is not normativist in import, for Stein intends it to apply to human beings specifically and not necessarily as a conceptual requirement. Hence, I omit further consideration of it.

[6] Cf. my defense (p. 62 above) of a strong (constitutive) reading of charity in Davidson over a weak (heuristic) one.

whether the textual evidence supports Davidson's adherence to a strong Competence Principle, or at least something closely approximating it.

Textual Evidence

In (2004b), Davidson endorses a view of irrationality that sees it as a kind of "inner inconsistency" on the part of an agent (189). Such a Humean conception is contemporary orthodoxy about what constitutes irrationality: it is a matter of a disharmony among one's own attitudes and actions, and does not hinge on assessment relative to facts and values external to the agent. But, more controversially, Davidson sees such inner inconsistency as, in some sense, involving a violation of the agent's own standards of thought and conduct. He writes, "It is only when beliefs are inconsistent with other beliefs according to principles held by the agent himself . . . that there is a clear case of irrationality" (2004b, 192). Moreover, what he takes to be true of beliefs, he takes to hold quite generally of attitudes and actions.

The dependence of irrationality on the individual, however, is attenuated by another facet of Davidson's account of irrationality. For Davidson holds that "all thinking creatures subscribe to *my* basic standards or norms of rationality" (2004b, 195). He maintains that a requirement for the possession of propositional attitudes at all is acceptance of "principles of decision theory", "the basic principles of logic, the principle of total evidence for inductive reasoning, or the analogous principle of continence" (2004b, 195). Now Davidson's formulations at this point may suggest that he is committed to a Competence Principle, where the relevant competence consists in a stock of beliefs whose content corresponds to basic normative rational principles. However, this reading is belied by Davidson's observation that when an agent has exhibited irrationality, "he must have departed from his own standards, that is, from his usual and best modes of thought and behavior" (2004b, 197). Davidson suggests that for him to "*have* the fundamental values of rationality" is to "show much consistency in his thought and action." In these passages, Davidson clearly equates one's "standards" to one's patterns of thought and behavior, not to one's normative beliefs.

Moreover, Davidson insists that in a case of irrationality, "the views, values, and *principles* that create the conflict are at that moment all active *tendencies* or forces" (2004b, 197). He maintains that "the elements that create the conflict" should not be viewed as "creating a *merely statistical preponderance* of the rational over the irrational." Rather, "all the beliefs, desires, intentions, and principles of the agent that create the inconsistency are present at once and are in some sense in operation—are live psychic forces." So Davidson flatly denies that the agent's necessary having the basic rational norms consists (merely) in their general conformity to the dictates of such principles.[7] Rather, one's having them is more a matter of exhibiting an active tendency to observe them. Thus, he seems to commit himself to the view that one's possession of them consists in *ceteris paribus* regularities corresponding to the principles holding true of one. Thus, with considerable plausibility, Davidson can be seen as subscribing to a Competence Principle, where one's rational competence—as on the mechanism view—embodies *ceteris paribus* regularities corresponding to (basic) rational norms.[8]

[7] Perhaps Davidson does commit himself to an agent's general observance of each individual basic rational norm. If so, this would amount to an additional charity principle of Davidson's. It is not one, however, that I wish to emphasize, if only because Davidson does not clearly advocate it.

[8] Is there reason to doubt that an agent can possess a normatively rational competence in the mechanism sense without possessing knowledge of normative principles too? The existence of physical devices like computers which can imitate (or perhaps instantiate) rational state-transitions

Clarification of a Few Points

A proper appreciation of this principle, however, requires clarification of a few points. First, any sort of mechanistic view may seem out of keeping with Davidson's trademark doctrine of the anomalousness of the mental.[9] But that doctrine specifically rejects only *strict* psychophysical (and psychopsychological) laws. There is nothing in Davidson's doctrine, or in his argument on its behalf , that would appear to commit him to the rejection of *ceteris paribus* laws of the sort bound up in the Competence Principle.

A second point concerns Davidson's apparent wish to distinguish between two kinds of principles of reasoning, basic and non-basic.[10] Evidently, he only means his Competence Principle to apply to such principles of rationality as are basic. However, in this regard two issues immediately arise, (1), just what distinction Davidson intends to capture with this terminology, and, (2), whether Davidson has the resources in his account of attribution needed to draw the relevant distinction. With respect to the first, Davidson cannot, of course, simply identify basic principles of rationality as those to which agents generally adhere. Rather, he needs to provide some handle on the relevant principles independent of their part in the Competence Principle itself if interpreters are going to be able to *apply* the Competence Principle in making attributions.[11] So just what *might* Davidson mean by "basic principles of rationality"?

without internal representations of the rules they observe suggests that a rational competence might well *not* require knowledge of normative principles.

[9] See (Davidson 1980c) for the classic statement of this doctrine.

[10] The passages from (Davidson 2004b, 195) quoted just above show this unmistakably. Cf. also Davidson's statement that the principle of total evidence is "so fundamental that we cannot make sense of an agent who does not generally reason in accord with it" (2004b, 190).

[11] Of course, if the number of such principle is small, it might be possible to identify basic principles simply by enumerating them. But Davidson provides no such list.

The interpretation might suggest itself that by ‘basic’ Davidson means ‘general’. The idea would be that basic rational principles are ones that subsume narrower, non-basic principles, rather in the way that a general logical principle like *modus ponens* subsumes a more special principle of the form, If *p* and *q* then *r*, *p* and *q*, therefore *r*. But if ‘basic’ is understood in this way, then an agent’s necessary disposition to obey the basic rational principles (enshrined in the Competence Principle) would guarantee their disposition to obey narrower, non-basic principles as well. However, Davidson pretty clearly does not intend the Competence Principle to guarantee obedience to non-basic principles. He intends a bifurcation between those rational principles adherence to which is conceptually mandatory and those adherence to which is not. So the interpretation in question, I think, should be rejected.

Davidson’s inclusion of “the basic principles of logic” (2004b, 195) in his representative list of principles that he takes to be fundamental in his sense might suggest a different line of interpretation. For in axiomatic and natural-deduction treatments of deductive logic at least, a distinction is often made between primitive and derived logical principles. Thus, in a natural-deduction system of the propositional calculus like E. J. Lemmon’s (1978), a small set of rules such as *modus ponens* and *modus tollens* are taken as primitive, and other rules, for example, what is often referred to as ‘disjunctive syllogism’, are justified in terms of them. However, this proposed interpretation is rendered problematic by the familiar fact that logic recognizes no single canonical set of primitive principles: the principles taken as primitive within one logistic may be regarded as derivative within another. As Wittgenstein taught us, considered in themselves, logical principles are all on a par:

no logical principle is really logically prior to any other (Wittgenstein 1961). Hence, despite its seeming promise, the primitive/derived distinction within logic is ill-suited to ground the basic/non-basic distinction that Davidson desires.

In fact, there appears no satisfactory non-psychological basis for the distinction. But a psychological basis would appear out of bounds for Davidson, or at least problematic. It might seem natural to view as basic those rational principles which, as a matter of empirical fact, agents universally possess a disposition to obey. But given the role that the basic/non-basic distinction plays for him as part of a constitutive constraint on the attitudes (that is, within the Competence Principle), such a view would obviously be circular.[12] Perhaps it would be open to Davidson to hold that interpreters possess an (*a priori*) theory of which rational principles agents are universally disposed to obey. However, this would greatly—and perhaps implausibly—increase the baggage with which Davidson would need to saddle interpreters. Moreover, Davidson's normativism would remain indeterminate until a list of such principles is provided; and it would impose upon the champion of this view the seemingly difficult task of motivating the inclusion of particular principles in the list. But perhaps these are not decisive objections. Accordingly, I shall proceed on the assumption that Davidson construes basic rational principles along the lines suggested.

Comparison to Other Charity Principles

A bit of comparison of Davidson's Competence Principle to his other charity principles is in order. First, the Threshold and Maximization Principles

[12] The same point would apply to any attempt to draw the distinction empirically, e.g., in terms of the obviousness or non-obviousness of violations of rational principles as assessed by agent's judgments of them.

fundamentally differ from the Competence Principle in that they constrain an agent's rational performance rather than their competence. Moreover—and relatedly—on the face of it the latter principle is logically independent of the former. It seems conceivable that an agent could possess a competence that includes all basic rational principles without exhibiting much actual rationality because subject to massive interferences with the exercise of the relevant dispositions.[13] Again, it appears *prima facie* conceivable that an agent's rational performance could be sterling but simply because they have had few occasions to engage their irrational competencies with respect to basic principles (or even because massive interferences have actually rectified the outcomes of those competencies!). So Davidson's introduction of a Competence Principle represents a substantive addition to his battery of normativist principles.

A significant feature of Davidson's Competence Principle is that in contrast to his other charity principles, it is non-holistic in character. Whereas the Threshold and Maximization Principles constitute tests applied to an agent's attitudes *en masse*, the Competence Principle is a test of each individual rational principle within an agent's rational competence. Consequently, the Competence Principle is vulnerable to straightforward refutation by counterexample in a way that its companions are not. Whereas the latter readily brook individual deviations from rational norms so long as a sufficiently high level of overall rationality is maintained, the Competence Principle bars even a single principle conflicting with basic rational ones from entering into

[13] Additionally, there is a question as to how much rationality would necessarily be secured even by perfect performance with respect to basic principles. For even in that case, non-basic principles might be massively violated, and there would appear no reason why these violations could not outweigh the exemplary performance with respect to basic principles.

one's rational competence. As will be seen, the argument of subsequent chapters exploits this vulnerability of the Competence Principle by making vivid the possibility of such deviant principles' being included in one's rational competence.[14]

It should also be mentioned that, although Davidson makes no explicit mention of this, presumably, he intends the scope of the Competence Principle to be epistemic as well as more properly rational. As I argued in Chapter Two, rightly understood, Davidsonian charity imposes on agents constraints as much concerning proper epistemic function as rationality. These sorts of normativity get, as it were, rolled up into one big ball in Davidson's account of the Threshold and Maximizing Principles. So I think it is no great stretch to interpret Davidson's Competence Principle as requiring agents to embody competencies reflecting basic epistemological norms.[15] At any rate, I shall proceed on the assumption that it does.

[14] There is some evidence that Daniel Dennett, like Davidson, may accept a variety of Competence Principle. Ed Stein, at least, views Dennett in this light (see Stein 1996, 116n14). Indeed, Dennett maintains that intentional systems "must be supposed to *follow* the *rules* of logic" if ascriptions of attitudes to them are to afford any "predictive power at all" (1971, 95). It is not altogether clear, however, whether Dennett means that such rules are embedded as dispositions in one's rational competence or merely that, as a statistical matter, agents must be supposed usually to observe them. But if Stein is right in attributing a view like Davidson's to Dennett, there remain important differences between them. Dennett seems to adhere to a *holistic* version of charity with respect to one's reasoning competence. For Dennett asserts, "not all the inference rules of an actual Intentional system may be valid . . ." (1971, 95). Nor is his point merely that such systems can dispense with non-basic principles. For Davidson's bifurcation of principles into basic and non-basic appears lacking in Dennett. In fact, Dennett even contemplates the possibility of agents who lack such a seemingly basic rule as *modus ponens* (at least in full generality) (1971, 95). Rather, Dennett's view seems to be that charity dictates merely that agents possess a preponderance of valid inference rules. In effect, he seems to advocate something like a Threshold Principle with respect to rational principles. (Cf. [Stein 1996, 124-27] on this brand of charity.) This has the significant implication that Dennett's version of normativism escapes that component of my argument of subsequent chapters that levels at Davidson's non-holistic Competence Principle. However, other components of the argument touch Dennett's version as much as Davidson's.

[15] In contrast to rational norms which appear to concern only propositional attitudes and actions and the relations among them, epistemological norms may treat of the proper relations of propositional attitudes (especially beliefs) to other sorts of cognitive mental states (e.g., sensations) as well (and perhaps also to the external world). On the reading I am proposing, then, Davidson holds that agents embody *ceteris paribus* regularities to link these items in ways corresponding to basic epistemological norms.

Argument for the Competence Principle

The interpretation of the Competence Principle aside, what reason does Davidson give us to believe it? Well, his argument (2004b, 195-96), though a bit obscure, on close inspection appears closely to co-incide with one recounted by Stephen Stich. Stich writes,

> It is part of what it is to be a belief with a given intentional characterization, part of the concept of such a belief, if you will, to interact with other beliefs in a rational way—a way which more or less mirrors the laws of logic. This sort of interaction with other beliefs is a conceptually necessary condition for being the belief that **not**-*p* or for being the belief that **if** *p*, **then** *q*. Thus if a belief fails to manifest the requisite interactions with other beliefs, it just does not count as the belief that not-p or the belief that **if** *p*, **then** *q*. (1990, 37)

Davidson, I think, essentially makes Stich's argument. However, his version concerns not just beliefs but propositional attitudes more widely. So it is not just "laws of logic" that must be manifested but principles of rationality generally—or, rather, those that are "basic." Moreover, one needs to be quite clear that manifesting "the requisite interactions" be taken as a matter of competence rather than performance if the argument is to be relevant specifically to establishing the Competence Principle.[16]

The problem with this 'argument' is that, in the most glaring fashion, it simply begs the question. There is the slightest conceptual distance between saying that to be an agent (a bearer of propositional attitudes) one's attitudes must be disposed to interact according to basic rational norms (the Competence Principle) and saying that to be a token-propositional attitude requires being disposed to interact according to basic rational norms. Anyone who harbors doubts about the former will harbor

[16] Although Davidson, of course, sees a high level of overall performance as requisite too.

doubts about the latter as well. So, I conclude, the Competence Principle remains unsupported.

Compartment Principle

According to the Competence Principle, divergences from basic rational principles must be viewed as performance errors, attributable to factors interfering with an inherently rational competence. This raises the question what interfering factors might give rise to such performance errors. At first sight, one would think that all sorts of factors, both purely physical and mental, might, in principle, do so. But, remarkably, there is some evidence that Davidson wishes to locate the source of such performance errors exclusively on the mental plane, more specifically, in the influence of one mental compartment upon another. So, ultimately, Davidson appears to envision a *refinement* of the Competence Principle along the following lines: A compartment of the mind embodies a capacity or competence for deploying propositional attitudes in accordance with basic standards of ideal rationality unless subjected to external interference by another mental compartment, resulting in performance error. When an agent exhibits irrational mental processes, then, Davidson sees this as due to the influence of one mental compartment or competence upon another (as opposed to, say, the influence of the agent's purely physico-chemical states). Let us call the resulting charity principle the *Compartment Principle*.

The Textual Context of the Principle

Davidson's most sustained treatment of mental compartmentalization appears in his "Paradoxes of Irrationality" (2004c). In this difficult article, Davidson's purpose is twofold: to resolve certain paradoxes associated with irrationality and, in the course of doing so, to defend certain key theses of Freudian psychoanalysis as conceptually required by an adequate account of (certain kinds of) irrationality. These include the claims, (1), that the mind includes semi-independent structures containing propositional attitudes and memories, (2), that the contents of parts of the mind can combine, as in intentional action, to cause events within the mind and without, and, (3), that some of the interactions among parts can be seen on the model of physical causation (2004c, 170-72).

The first paradox Davidson identifies springs from the way we describe and explain propositional attitudes. An element of rationality seems to be built into our very descriptions of propositional attitudes and the fact that we explain them in rationalizing terms. For example, to use Davidson's illustration, Roger's intention "to pass an examination by memorizing the Koran . . . must be explained by his desire to pass the examination and his belief that by memorizing the Koran he will enhance his chances of passing the examination" (2004c, 169). Since fitting attitudes "into a rational pattern" through rationalizing explanations is inseparable from the attitudes, it can seem puzzling how irrationality can exist at all (2004c, 169-70).

This first paradox is, I think, easily dissolved. In the first place, Davidson's claim that the very descriptions of the attitudes implicates them in rationalizing explanation is made to seem plausible only by his cherry-picking examples, such as Roger's complex instrumental intention. A description of a simple attitude like the belief that the earth is round, by contrast, in no way seems necessarily to call for a rationalizing explanation. Moreover, even if rationalizing explanation *were* inseparable from the attitudes, one easily sees how irrationality would be still be possible. For a demand for explanation by reasons would not be tantamount to a demand for explanation by *good* reasons. A belief, for example, can be held for a reason (i.e., on the basis of other beliefs one holds) but still count as irrational because fallaciously inferred from those other beliefs.

Davidson's manner of dissolving this paradox, however, is rather different, perhaps because of his focus on certain kinds of irrationality in this article, namely, *akrasia* and wishful thinking. Adapting a case considered by Freud, Davidson describes an akratic man who while walking in the park, removes a branch from his path , but later, thinking it a danger to passersby in the hedge into which he has tossed it, returns to the park to replace it, even though he realizes that he has motives not to return—the time and trouble involved—which outweigh his concern for the safety of passersby (2004c, 172-74). Moreover, Davidson describes a case of wishful thinking where a "young man very much wishes he had a well-turned calf and this leads him to believe he has a well-turned calf," where "the entire explanation of his holding the belief is that he wanted to believe it" (2004c, 178). Davidson asserts that in such cases of irrationality there is a mental state that causes the relevant action or propositional attitude without serving as a reason for it. Thus, in the case of wishful thinking, "a desire causes a belief. But the judgment that a state of affairs is, or would be, desirable, is not a reason to believe that it exists" (2004c, 179).[17]

This introduction of non-reason mental causes into the account of irrationality brings in its train an appeal to mental compartmentalization. For rather inscrutable reasons, Davidson finds such causes problematic. Apparently, the only way

[17] There might be some room for debate about this. Granted, a desire is not a theoretical reason for a belief, but the view that there can be practical reasons for beliefs is not without its defenders.

Davidson can conceive of such non-reason mental causation is when cause and effect are segregated in different minds, or at least different mental compartments. As Davidson notes, non-reason mental causation between minds appears unproblematic. Thus, for example, my wish for you to enter my garden may lead me to "grow a beautiful flower there," which may, in turn, entice you to enter (2004c, 181). Moreover, what applies in the case of such social interactions, Davidson holds, can apply to a single person as well. Indeed, Davidson writes, "if we are going to explain irrationality at all, it seems we must assume that the mind can be partitioned into quasi-independent structures that interact . . . " (2004c, 181).

Thus, in (2004c) Davidson seems to have arrived at a Compartment Principle—or something close to one: an agent's irrational mental processes are due to the influence of one mental compartment upon another. So Davidson takes himself by this route to have vindicated, among other things, the coherence of Freud's appeal to mental compartments. Far from incoherent, such compartments are required to explain (a form of) irrationality.[18]

[18] A point that is easily missed is that the influence of one compartment upon another as required by the Compartment Principle in instances of irrationality is not to be conceived of as direct intentional action. In a case of self-deception, for example, the intention in Compartment$_1$ that there should be a belief that *p* in Compartment$_2$ does not cause that belief directly in the manner of basic actions, as my intention to raise my arm causes my arm to raise. Rather, as in Davidson's analogy of enticing someone into one's garden with a beautiful flower, the intention produces the effect indirectly: "What is essential [to the analogy] is that certain thoughts and feelings of the person be conceived as interacting to produce consequences on the principles of intentional actions, these consequences then serving as causes, but not reasons, for further mental events" (2004c, 185). Thus, the picture is the following: Compartment$_1$acts so as to produce consequences *C* which, in turn, produce a belief, say, in Compartment$_2$.

There is much puzzling about this picture, however. First, one might wonder why Davidson should insist upon it, as opposed to direct action. If the interaction were direct, though, perhaps one would be tempted to say that the intention in Compartment$_1$ is, in fact, a reason for the belief in Compartment$_2$, which runs counter to Davidson's insistence that instances of irrationality involve non-reason mental causes. Or would such direct interaction be inconsistent with Compartment$_1$ and Compartment$_2$ being separate compartments? In any case, it is intuitively odd to imagine a mental compartment conniving through means-end reasoning to produce physical consequences in order, ultimately, to induce an effect in some other mental compartment! Davidson, however, writes that we should not "speak of

Unclear, however, is whether Davidson intends the compartmental model to apply to *all* irrationality. In (2004c), Davidson stops short of committing to this. In the later article, (2004b), however, he comes closer to an unqualified assertion of the Compartment Principle. There he writes, "What is needed to explain irrationality is a mental cause of an attitude, but where the cause is not a reason for the attitude it explains" (2004b, 190); and he appears to assert quite generally, "it is only by postulating a kind of compartmentalization of the mind that we can understand, and begin to explain, irrationality" (2004b, 198). At all events, in assessing Davidsonian charity, I feel the textual evidence justifies considering the plausibility of the Compartment Principle, along with that of other charity principles to which Davidson subscribes.

Critical Discussion

But Davidson's derivation of the Compartment Principle itself requires some critical discussion. An appeal to non-reason mental causes is problematic for at least two reasons. First, it is not obvious how much irrationality they are implicated in. Davidson cites wishful thinking and *akrasia* as involving such causes (see p. 114 above), and there is some plausibility in regarding wishful thinking, perhaps even as a matter of conceptual necessity, as involving the non-rationalizing causal influence of a wish in the formation of a belief. That non-reason mental causes are implicated in *akrasia*, however, is far from obvious. Granted, Davidson's own account of *akrasia* finds a place for such causes, in the influence of the mental states of one compartment

parts of the mind as independent agents" (2004c, 185). It is unclear, though, how Davidson's picture avoids characterizing parts as agents. Perhaps despite appearances, Davidson should actually be taken as holding that the intentions of Compartment$_1$ do not really produce consequences *C* "on the principles of intentional action" but, rather, merely as ordinary causal consequences, which, in turn produce mental states in Compartment$_2$. This may be the most charitable reading.

upon another. But there appears nothing in the very concept of *akrasia* that dictates such an account. Indeed, other philosophers provide accounts that dispense with any necessary role for non-reason mental causes.[19] Of course, even less obvious is that forms of irrationality consisting of making inferences according to patterns that deviate from canonical logical principles must involve non-reason mental causes. On their face, they seem to involve causes that are reasons—ones, however, that just happen to be bad reasons. Moreover, there is the difficulty of seeing why non-reason mental causes must operate *across* mental compartments and not within them. As noted, Davidson's suggestions to that effect are obscure. Without clear and compelling argument, there seems little reason to see Davidson's proposal that

[19] George Rey's account of *akrasia*, for example, is based on a distinction he draws between central and avowed attitudes (Rey 1988). According to Rey, one possesses two distinct sets of attitudes, one's central beliefs and preferences, which "enter into instances of practical reasoning which largely determine one's acts," and one's avowed beliefs and preferences, which underlie one's sincere assertions (Rey 1997, 294). It is the capacity for discrepancies to arise between these two sets of attitudes that is foundational for Rey's accounts of *akrasia*.

Rey gives the following concise description of *akrasia*: "*Akrasia* occurs when someone avows all the relevant preferences, avows one to be higher than the other; but still centrally values the other over the one" (1988, 282). The akratic's situation, then, appears to be this: she has avowed preferences P1 and P2, and an avowed belief that P1 is of greater weight than P2. Moreover, she simultaneously has central preferences P1* and P2*, with the same content as P1 and P2, respectively, but such that P2* is of greater weight than P1*.

An example will make things more vivid. Suppose an individual confronts a choice at a social function between either sticking to his vegetarian principles (but going hungry) or giving in to his hunger and partaking of a meal containing meat. Avowedly seeing reasons supporting either option, he may avowedly believe that on balance the reasons which support sticking to his principles are greater. Yet he may find himself choosing to eat because he centrally assigns greater weight of reasons to that option.

What is of interest in Rey's account in the present context is that it dispenses with non-reason mental causes in explaining akratic acts. On his account, such acts are completely explained by the central beliefs and preferences which rationalize them. One might object that little is gained by way of criticizing Davidson's attempted derivation of a Compartment Principle by this observation, since Rey himself finds it necessary to appeal to a kind of compartmentalization to account for *akrasia*, namely, that among avowed and central attitudes. On closer inspection, however, his account of *akrasia* is seen to run counter to the Compartment Principle. For though it adverts to compartments, it makes no appeal to the interfering influence of one compartment on the other, which the Compartment Principle demands in cases of irrationality.

irrationality involves compartmental interaction as more than an empirical hypothesis, not as the conceptual requirement that Davidson plainly takes it to be.

However, Davidson does not posit compartments solely in order to accommodate the supposed element of non-reason mental causation in irrationality. At points, he lays stress on how irrationality can involve a subject's entertaining logically inconsistent beliefs. Thus, in "Deception and Division," Davidson's account of self-deception involves attributing to a subject simultaneously the belief that *p* and the belief that not *p* (2004a, 208). But Davidson holds that such inconsistent beliefs must belong to distinct mental compartments: "we must accept the idea that there can be boundaries between parts of the mind; I postulate such a boundary somewhere between any (obviously) conflicting beliefs" (2004a, 211).

Note, however, that even if such boundaries are necessary in cases of crass inconsistency, this is still not enough to yield a Compartment Principle, since there is no suggestion that all irrationality involves such crass inconsistencies.[20] So this alternate route to compartments also fails to yield a Compartment Principle.[21] An additional difficulty with Davidson's proposal—of a sort familiar by now—is that it relies on making a bifurcation between obvious and non-obvious inconsistencies. Davidson plainly does not intend it to apply only to beliefs whose contents are in formal contradiction, so he owes an account of which beliefs are obviously

[20] The premises might also be questioned that it is an *a priori* fact that the belief that *p* and the belief that not *p* cannot belong to the same mental compartment.

[21] It is worth noting that self-deception, to the extent that it involves crass inconsistency, constitutes a form of irrationality that also involves non-reason mental causes. For as Davidson points out, in central cases of self-deception the belief that not *p* is itself causally implicated in engendering the very belief that *p*. But, of course, the fact that not *p* is not a *reason* to believe its contradictory *p* (2004a, 208-09).

inconsistent. But it is implausible that he can draw this distinction with the meager resources of his (apparently) purely *normative*-based interpretativism.[22]

It is also worth noting ways in which the appeal to compartments may ultimately be problematic for normativists. In the first place, it is by no means clear that Davidson's detailed account of attribution (see p. 64ff. above) applies without modification to a compartmentalized mind, or if not, whether and how Davidson can modify it so that it does. Central to Davidson's account is that it intimately weds attribution of an agent's propositional attitudes with interpretation of their language. What Davidson offers is "a combined theory of meaning and belief" and desire (2004e, 156), on which attribution depends on observable evidence about whether an "agent prefers one sentence true rather than another" (2004e, 158). It is only to the extent that attitudes rationally and causally condition such preferences (and their overt expression) that renders them susceptible of attribution.[23] Hence, attitudes which have no direct rational and causal bearing on an agent's verbal behavior seem to fall out of the scope of Davidson's account. In essence, then, it *appears* that Davidson's account cannot straightforwardly accommodate unconscious attitudes. But, leaving to one side such phenomena as Multiple Personality Disorder, compartments of the sort Davidson contemplates would be aphasic, and their contents unconscious. Davidson's account seems unable to ground attribution of such contents. Now this difficulty might seem to lapse where, as in Freudian theory,

[22] Cf. the discussion above (p. 106) concerning the difficulty Davidson confronts in distinguishing between basic and non-basic principles of rationality.

[23] Thus, Davidson writes, ". . . the preferring true of sentences by an agent is . . . clearly a function of what the agent takes the sentences to mean, the value he sets on various possible or actual states of the world, and the probability he attaches to those states contingent on the truth of the relevant sentences. So it is not absurd to think that all three attitudes of the agent can be constructed on the basis of the agent's preferences among sentences" (2004e, 158-59).

contents are—at least in significant measure—*capable* of becoming conscious. To the extent these contents become conscious, they will typically acquire the sort of connection to linguistic behavior that on Davidson's account permits their attribution. However, it will be noted, to the extent they become conscious, they are no longer to be reckoned to unconscious compartments. At best, Davidson's account seems to allow their attribution *qua* conscious. While they are unconscious, for all Davidson's account is concerned, it is as if they do not exist. So Davidson's account, if it is to accommodate compartments of the very sort Davidson himself wishes to allow for, appears in need of substantial supplementation.

In the second place: As Davidson observes, certain forms of irrationality seem to involve attributing crass inconsistencies to an agent. Davidson, like many non-normativists, reaches for the expedient in such cases of quarantining inconsistent beliefs in separate compartments. However, if the notion of contradictory beliefs within a single compartment offends against Davidson's normativist intuitions, it is far from obvious that the introduction of multiple mental compartments to accommodate them readily comports with those same intuitions either. Now in light of the introduction of compartments, one might think that Davidson intends his earlier enunciated charity principles (such as the Maximization and Threshold Principles) to be applied to compartments individually: compartments should be treated as isolated spheres within which individually truth and coherence should be maximized, etc. However, that is not Davidson's view. Maximization and Threshold Principles remain in effect as global constraints on the mind as a whole.[24] But as Davidson

[24] By the same token, it is presumably Davidson's intention that the Competence Principle also be thought of as applying to the entire cognitive system. Thus, to fulfill this requirement, it will not

himself notes, the Maximization Principle enjoins seeking a unified, consistent interpretation of an agent insofar as possible. So to the extent that partitioning accommodates inconsistences, it might seem to run counter to charity's imperative to minimize them (2004c, 182-84).

The potential conflict between compartmentalization and charity is particularly keenly grasped by Christopher Cherniak. As I pointed out earlier, Cherniak and Davidson share a broadly normativist outlook, and they agree in according a place to compartmentalization in their pictures of the mind. But whereas the route Davidson takes to acknowledging mental compartments is through the existence of irrational phenomena, Cherniak posits them on the basis of more mundane facts about the structure of memory. Cherniak sees it as a non-contingent feature of our commonsense picture of the mind that it contains distinct memory storages, along the lines suggested by cognitive-psychological models which distinguish between short-term and long-term memory (1986, 54-56). Moreover, that commonsense picture further compartmentalizes long-term memory in that it "assumes an organization of long-term memory, one that determines the pattern of a search for an item . . ." (1986, 56).[25] Cherniak holds, further, that once this picture is taken into account, one will not be tempted—as Quine clearly was in formulating his principle of charity—to deny that minds can entertain obvious logical inconsistencies (1986, 56). Such compartmentalization makes intelligible the possibility of abundant

suffice for individual compartments to appear to embody rational competences only when viewed in isolation. Rather, they must be seen to do so against the backdrop of the entire cognitive system of which they are a part. Generally, my arguments against normativism in succeeding chapters will be directed against versions which, like Davidson's, take the whole cognitive system, not individual compartments, as the unit to which charity standards apply (*monocentric* normativism). However, at points I consider the prospects for arguing against a *polycentric* normativism.

[25] Cherniak notes that cognitive-psychological theory contemplates the possibility that both 'working' (i.e., short-term) memory and long-term memory may, in fact, be multiple (1986, 53).

inconsistency: “Logical relations between beliefs in different ‘compartments’ are less likely to be recognized than relations among beliefs within one compartment, because in the former case the relevant beliefs are less likely to be contemporaneously activated, and . . . it is only when they are activated together that such relations can be determined” (1986, 67).[26]

But, like Davidson, Cherniak recognizes a tension between the possibility of abundant inconsistency that compartmentalization raises and holistic charity constraints. Essentially, the problem is that such inconsistency can potentially clash with the Threshold Principle.[27] He writes,

> The cost of compartmentalization is some isolation of subsets of the belief system from each other, and the resulting lack of interaction can fragment the total system. The contents of long-term memory are subject to less stringent rationality requirements than the contents of short-term memory, but they are not permitted unlimited irrationality. Only a balance of compartmentalization of long-term memory enables a complete cognitive system to qualify as minimally rational. (1986, 69)[28]

Too much compartmentalization, because of the attendant inconsistency and irrationality that it introduces, must be ruled out. This potential clash between compartmentalization and charity will loom large in my argument of the next chapter.

[26] Worth noting is that, unlike Davidson, Cherniak apparently does not rule out the possibility of obvious consistency within a single compartment. Rather, he merely stresses how compartmentalization increases the likelihood of inconsistencies.

[27] Although, of course, Cherniak, with his moderate normativism, sets the threshold of requisite rationality lower than does Davidson.

[28] In—and around—the quoted passage, Cherniak is concerned to make at least two distinct, though related, points: (1) that too much compartmentalization tends to conflict with charity requirements, and (2) too much compartmentalization threatens the degree of integration required for personhood. The latter point concerns the possibilities of causal interactions among one’s propositional attitudes, rather than their logical relations *per se*. Accordingly, I think it more properly falls under the heading of humanity than that of charity.

Mental Modularity

In general, my argument of the next chapter will suggest that, in various ways, mental compartments, specifically, in the form of psychological *modules*, pose problems for Davidsonian charity. So it is appropriate that I set out the relevant sense of ‘module’ on which my argument relies and clarify how such modules relate to the compartments that Davidson (and Cherniak) discuss.

The Notion of a Module

With respect to the divisibility of mind, the philosophical tradition has itself been somewhat ‘schizophrenic’. On the one hand, philosophers like Plato and Descartes have often insisted on the simplicity of the mind, especially when doing so has seemed a way to establish the immortality of the soul. But, on the other hand, in the *Phaedrus* and elsewhere Plato himself famously presents a tripartite division of the mind or soul into rational, spirited, and impulsive elements (Plato 1961). Psychologists, by contrast, in recent years at least, have been fairly ‘unanimous' in recognizing the mind’s divisibility on one or another basis. A number of pathological phenomena—including multiple personality disorder, split-brain phenomena, and blindsight, to name a few—seem to call into question the picture of the mind as a unified sphere of consciousness, at least in abnormal cases. But psychologists have found it expedient to see the normal mind too as consisting of parts, and many of their specific proposals in recent years have employed the notion of a module.

Gabriel Segal (1996) provides a taxonomy of various conceptions of modules present in the literature which, he believes, “have a good chance of being genuine psychological natural kinds” (1996, 141); and preparatory to defending a thesis of “massive mental modularity”—the view that “the mind consists entirely of distinct components, each of which has some specific job to do in the functioning of the whole" (Carruther 2006, 2)—Peter Carruthers analyzes current conceptions of modules in great detail. Of course, to the extent that the mind is universally acknowledged to possess various faculties, there is a sense in which it is uncontroversial that the mind consists of parts. But the notion of a module, as both Segal and Carruthers make clear, is more robust than that of a mere faculty. Rather, they are theoretical entities posited in order to account for psychological faculties or capacities (Segal 1996, 141). Moreover, there seems agreement that individual representations (or arbitrary sets of such representations) do not count as modules (cf. Carruthers 1996, 3).

Segal, in his classification, makes a basic distinction between synchronic and diachronic modules. But since diachronic modules in essence merely seem to amount to a kind of higher-order module—modules for the acquisition over time of synchronic modules—I shall focus mainly on the varieties of synchronic modules, that is, modules that underly capacities that one can exercise at some particular time.[29] Among these, Segal identifies the following kinds: (1) intentional modules, (2) computational modules, (3) Fodor modules, and (4) neural modules. Segal describes an intentional module as "a specific body of psychological states" that underlies a competence and cites as examples the Freudian unconscious[30] and

[29] Segal's characterizes a synchronic module generally as "a component of the mind, or brain, a mechanism, a system or some such that explains [a] competence" (1996, 142).

[30] Smith (1999, 120) distinguishes different senses in which Freud uses 'unconscious':

> . . . Freud uses the term 'unconscious' in several ways. Sometimes the term is used to designate a functional system of the mind containing mental representations that are unconscious but not preconscious (and possessing special irrational characteristics . . .). . . . At other times Freud uses 'unconscious' to denote all of those mental items that are not

Chomsky's view that our linguistic competence consists in part in our tacit knowledge of linguistic rules (1996, 143).

But the stark differences between Freudian unconscious and Chomskyan linguistic competence suggest that two sorts of things are perhaps run together in Segal's notion of an intentional module.[31] Accordingly, it may be worthwhile to attempt to achieve some clarity about what those two sorts of things might be. An issue is what bodies of psychological states are meant to count as intentional modules. Segal hints that they need to be "appropriately related" in some way, but he is vague as to what relation is constitutive. With respect to Chomsky, Segal stresses the thematic, conceptual relations shared by the items of linguistic knowledge that for Chomsky underly our linguistic competence: "The knowledge concerns a self-contained array of interrelated concepts (Phrase, Noun, Verb, Anaphor, Quantifier, etc.) that fit together some what in the manner of a scientific theory . . ." (1996, 143).

But Segal is aware that some other basis of unity needs to be employed to effectively carve out the notion of intentional module towards which he is striving. He maintains that "Mere knowledge of a theory isn't likely to be a psychologically interesting category" (1996, 143). That is, the sort of thematic coherence characteristic of a theory (as holds in the case of Chomskyan linguistic competence) is not sufficient for being an intentional module. Accordingly, using Fodor's

> consciously represented . . . . Finally, Freud sometimes describes as 'unconscious' all mental items under neuroscientific description . . . .

Apparently, it is something like the first sense that Segal has in mind in maintaining that the Freudian unconscious counts as an intentional module. My argument of subsequent chapters, however, will involve describing modules that, though working with states belonging to the Freudian unconscious, in fact, constitute what I subsequently term 'processing modules' (see p. 127 below).

[31] In a footnote, Segal acknowledges that he provides "only a very quick first sketch of intentional modularity" and that "If one or more genuine psychological natural kinds fall under the concept then a lot of work remains to be done articulating it" (1996, 157n2).

terminology, Segal attempts to fill the breach with the requirement that the relevant body of psychological states exhibit either ‘informational encapsulation’ or ‘limited accessibility’ vis-à-vis the contents of rest of the mind (but especially consciousness). Thus, for example, Chomsky’s tacit linguistic theory is inaccessible to consciousness. Segal, then, stipulates that an intentional module is “a set of appropriately related psychological states” which “exhibits either informational encapsulation or limited accessibility” (1996, 143).

But the only sort of appropriate relation Segal has broached is the thematic one. This, however, is lacking in the Freudian case, for Freudian theory permits quite various contents to be consigned to the unconscious. So all Segal really has to hold together the category of intentional modules is the notion of an encapsulated or inaccessible set of psychological states. Does this suffice to capture an interesting category of module? I don’t know. But, in any case, the differences between Freudian unconscious and Chomskyean linguistic competence seem sufficiently great to—at the very least—represent instances of distinct species of module, whether or not they fall under a common genus which Segal would designate ‘intentional module’.

Thus, Carruthers cites R. Samuels (1998) as a proponent of what the latter terms ‘informational modules’. These are “organized bodies of innate information” in specific domains which can be drawn upon as the mind performs the tasks related to given domains (Carruthers 2006, 32). The sort of intentional module represented by Chomskyean linguistic competence appears in all essentials to coincide with Samuels’informational modules. The Freudian unconscious, by contrast, would clearly not count as an informational module. Of course, there is the obvious point that the Freudian unconscious contains affective and desiderative mental states besides ones aptly called ‘informational’. But the lack of thematic unity, again, is the decisive difference. Accordingly, I propose to call the species of intentional module represented by the Freudian unconscious ‘a non-thematic module’. There would seem to be nothing particularly abstruse about such modules. Indeed, any sort of memory storage that can receive disparate contents would also count as such.

Both informational and non-thematic modules contrast with Segal’s category of ‘computational modules’. These are processors which perform some specific mental function by operating on physical representations in a language of thought (1996, 143-45). So, for example, psychological theorists have posited computational modules corresponding to such mental functions as mind-reading, language-processing, and reasoning in various domains. As the name makes plain, the conception presupposes the computationalist view of the metaphysics of mind, which both Segal and Carruthers are concerned to defend (cf. Segal 1996, 148-49; Carruthers 2006). But Carruthers is careful to allow the possibility of processors which operate along non-computational, connectionist lines (2006, 45n23). Hence, it seems sensible to acknowledge a genus of processing modules under which one can distinguish both computational and non-computational modules as species.

Segal identifies Fodor modules as a species of computational module, one that possesses a number of specific properties: “(1) Domain specificity (2) Informational encapsulation (3) Obligatory firing (4) Fast speed (5) Shallow outputs (6) Limited inaccessibility [*sic*] (7) Characteristic ontogeny (8) Dedicated neural architecture (9)

Characteristic patterns of breakdown" (1996, 145). Fodor's inclusion of several of these properties (such as fast speed and shallow outputs) is explained by his exclusive focus on input-output systems like the early visual system. But Segal envisions the possibility of non-Fodorian computational modules which dispense with some of these properties. Moreover, Carruthers, in the context of defending his thesis of "massive mental modularity," explicitly develops a notion of computational module considerably weaker than that of a Fodorian module. He argues that if that thesis is to be remotely plausible several of the Fodorian requirements must be excluded. Thus, for example, if modules are to account, not just for perception, but for central processes of belief-fixation, desire-formation, and planning, then modules must be allowed to deliver deep, conceptual outputs, not merely shallow ones.

There remains the neural module, which Segal describes as "a functional component of the brain, describable in purely neurological terms" that subserves some particular cognitive capacity (1996, 145). Discussions of modularity take place fairly universally against the background of an assumption of physicalism. So proponents of modules will hold that modules are neurally realized in some fashion.[32] Neural modularity, however, is the view that they can be mapped onto specific neurological systems, where those systems are individuated in terms proper to neurophysiology, instead of being realized by global features of the brain. Although the issue whether modules are realized by neural modules is an important empirical

[32] It is worth noting, however, that there seems no obvious conceptual bar to entertaining the possibility of modules, both intentional and processing, that are non-physical.

question, it is a bit remote from the—more conceptual—concerns of the present project. So I shall ignore the issue of neural modularity in the sequel.[33]

My discussion, then, yields the following classification of types of modules:

1. Processing
    1.1 Computational
        1.1.2 Fodorian
        1.1.3 Non-Fodorian
    1.2 Non-computational
2. Non-Processing (‘intentional’)
    2.1 Thematic (‘informational’)
    2.2 Non-thematic

Davidsonian Compartments as Modules

Where, if anywhere, might Davidson’s compartments fit in this classification? Well, as Davidson himself notes, he characterizes compartments at a highly abstract level: “In particular,” he writes, he “has nothing to say about the number or nature of divisions of the mind, their permanence or aetiology” (2004c, 186n6). Nonetheless, his account is not so indefinite as to allow some conclusions to be drawn. First, his compartments have at least a *connection* with with intentional modules, specific bodies of psychological states[34] that exhibit informational encapsulation (and limited accessibility). Indeed, their very raison d΄être within Davidson’s account of

[33] As noted above, I shall also ignore the diachronic, module-forming modules. But Chomsky’s LAD, short for ‘Language Acquisition Device’, can serve as an example of such. This is a postulated innate system which in response to environmental stimuli gradually issues in a developed linguistic competence (for Chomsky, an intentional module). In general, the synchronic modules which a diachronic module outputs can be either processing modules or purely intentional ones.

[34] Not that these bodies of psychological states should be thought of as mere *sets* of states, constituted extensionally. For Davidson plainly allows compartments to persist over time despite changing their contents. (Indeed, the same holds for all varieties of modules. Even the contents comprised by informational modules can in principle be acquired gradually, grow, and otherwise develop.)

irrationality is to allow for the segregation (that is, the causal-functional isolation) of mental contents from other mental contents.[35]

But it would be quite a stretch to *identify* Davidson's compartments with intentional modules. Davidson plainly sees them as containing, not just beliefs, but desires and intentions as well. Moreover, Davidson's compartments are not mere collections of information (or other mental states), as are intentional modules. Rather, among other things, they are meant to be arenas in which mental processes take place, in which beliefs and desires "can combine, as in intentional action, to cause further events in the mind or outside it" (2004c, 171). So, to this extent, they appear to amount to a kind of processing module.[36] But processing modularity, as it appears in Fodor and Carruthers, for example, appears closely tied to the execution of particular cognitive functions: a processing module is a component of the mind that "has some specific job to do in the functioning of the whole" (Carruthers 2006, 2), whether it be mind-reading, language-processing, or reasoning[37] However, this tie to particular functions is absent from Davidson's characterization of his compartments.

[35] There are limits, however, to the amount of encapsulation that Davidson is prepared to permit. Given Davidson's content-holism, he is committed to holding that to the extent that different compartments contain mental states whose contents share conceptual components, those compartments must significantly overlap. Thus, he writes, "We should not necessarily think of the boundaries [between compartments] as defining permanent and separate territories. Contradictory beliefs about passing a test must each belong to a vast and identical network of beliefs about tests and related matters if they are to be contradictory" (2004a, 211). But this does not impact the tie between Davidsonian compartments and intentional modules, for Segal's notion allows for partial encapsulation (or inaccessibility).

[36] It should be noted that the mere fact that Davidson's compartments appear agentive in character would not preclude them from counting as processing modules. Carruthers (2006), e.g., in his envisioned architecture postulates practical-reasoning modules which, taking beliefs and desires as inputs, issue intentions as outputs.

[37] For those like Carruthers who embed modularity in the context of evolutionary psychology, the function of a module would largely seem to be determined by its contribution to biological fitness. However, Carruthers and others recognize a category of learned modules (contrasting with innate ones) which are acquired in the course of learning particular skills, e.g., motor skills (2006, 10n6). The functions of such modules, whose possession can be highly contingent and variable, might, it appears, need to be identified along other than evolutionary lines.

Granted, as a class they serve an explanatory role for him in accounting for irrationality, but it would be odd to regard this as a *function*; and even if it were, Davidson by no means limits their operation so narrowly. In fact, for all Davidson says about the matter, they might exhibit just about *any* cognitive function or functions.[38] Again, Davidson at points seems somewhat inclined to regard his compartments as temporary. Though, as noted, in (2004c) Davidson remains neutral about whether they are permanent or not, in (2004a) he asserts, “We should not necessarily think of the boundaries [between inconsistent beliefs] as defining permanent and separate territories” (211). Whatever his reason for holding this, if the compartments *are* temporary this would be a respect in which they differ from the relatively permanent sort of processing module that Fodor and Carruthers envision. Nonetheless, in terms of the typology set out above, overall they most closely resemble processing modules.

General Structure of My Argument

Having clarified the notion of modularity that will figure in my argument against normativism, I now set forth the general structure of that argument. The aim is to envision minds that do not hew to Davidson’s Competence, Threshold, or Compartment Principles. So I describe parts of minds, modules, that evidently embody irrational competences, thereby conflicting with the Competence Principle. Moreover, since these modules diverge from standards of ideal rationality in virtue of their normal operations and not through the external influence of some other mental

[38] Indeed, as Davidson describes them, they appear rather homuncular. It seems that potentially they might incorporate all or most of the functionality of the whole person. In this regard, a comparison of his compartments to the distinct personalities involved in Multiple Personality Disorder might not be inapt, at least if those personalities be thought of as subsisting simultaneously.

compartment, they also violate the Compartment Principle. But, further, these modules (or collections of them) can be envisioned as subsisting in a mind prior to its full development, either ontogenetically or phylogenetically, in such a way as to largely exhaust its capacities for propositional thought and reasoning. In such a mind, irrational processes would clearly predominate and, therefore, the mind itself would violate the Threshold Principle.[39] The conceivability—indeed the scientific possibility—of the hypotheses I adduce, then, suggests the untenability of those principles.

Forestalling Objections

Before setting forth my argument in detail, however, I wish to forestall certain objections that might suggest themselves at the outset. Specifically, for various reasons, one might argue that the sorts of modules I shall describe escape the intended sphere of application of normativist constraints and that, therefore, they cannot be legitimately adduced as violations of those constraints. That is, a type of counter-

[39] Some word of explanation of why the argument is couched specifically in terms of modules are in order. For in principle there would seem no (obvious) conceptual bar to envisioning a competence lodged, not in a module, but in the more holistic function of a mind. If that competence were irrational, it would violate the Competence Principle, of course, and the Compartment Principle (because if the divergence from rationality is not due to performance error *a fortiori* it is not due to performance error induced by an interfering compartment). But for several reasons it appears preferable to make the argument specifically in terms of modules. First, as a matter of empirical fact, it is more likely that such *ceteris paribus* regularities as constitute psychological competences will be found in subsystems of the mind, like modules, than in the mind's more holistic operation. As Georges Rey observes—with respect to intentional action, but the point generalizes—"Ordinary human behavior is arguably the result of interaction among a multitude of probably quasi-independent subsystems of the mind . . . and where there is such complex interaction, one seldom expects there to be any clear, scientifically respectable *laws*, at least not at the level of the interaction. The laws concern regularities about the *subsystems* . . ." (2001, 111). Second, and relatedly, the scientific hypotheses about irrational competences—or at least the most plausible versions of them—which lend themselves to employment in my argument are formulated in terms of parts of the mind, in terms of modules. Third, although irrational competences lodged in holistic mental function could be used to argue against Competence and Compartment Principles, they would not serve as readily to argue against the Threshold Principle. For that argument depends on the detachability of competences, a feature that competences based in modules can possess, whereas ones based in holistic mental function cannot. The dissociability that *can* attach to modules forms a key element, then, in my argument against the Threshold Principle.

argument can be mounted which has the following form: Normativists constraints apply only to mental processes which possess property *P*; mental processes within modules do not possess *P*; therefore, normativist constraints do not apply to them. There are several properties which with some initial plausibility might be selected for substitution into this schema: It might be held that normativist constraints apply only to mental processes with propositional content (propositional attitudes); or only to certain sorts of propositional attitudes (specifically, decision-theoretical ones); or, again, only to mental processes which are rationality-evaluable. Moreover, different reasons can be advanced for thinking that the processes in modules lack the relevant properties, for example, because they are ‘subdoxastic’ or ‘subpersonal’ or non-inferential. In the present section, I shall canvass the various considerations that speak for and against the contentions underlying these arguments. My response will be to suggest that, in each case, these arguments either construe the scope of normativism too narrowly or else exclude modular processes from that scope on flimsy grounds.

An Argument Based on a Distinction from Stich

One potential argument is founded on a contention, associated with Stephen Stich (Stich, 1978), that tacit mental processes of the sort liberally postulated in recent information-processing cognitive psychology do not involve genuine beliefs but rather distinctive ‘subdoxastic’ states, which, in fact, on one reading lack propositional content altogether. Since the states involved in modular hypotheses like those I consider in later chapters are of the sort Stich would reckon subdoxastic,[40] an argument against my employment of these modules against normativism could be

[40] A complication here is that Stich actually sees the states postulated by Freudian theory as doxastic. But this is largely because he mistakes the character of such states (see p. 137, n. 46 below).

constructed along the following lines: Normativist constraints apply only to beliefs (and other decision-theoretical attitudes); modular mental states are not beliefs (or other decision-theoretic attitudes); therefore, normativist constraints do not apply to modules.[41] The argument merits close scrutiny. First I examine Stich's contention, then the premise that limits normativist constraints to decision-theoretic attitudes.

Beliefs versus Sub-Doxastic States

Stich makes a distinction between beliefs and 'sub-doxastic states' which he understands to be "non-belief states that play a role in the proximate causal history of beliefs" (1978, 499). Regarding this as an intuitive distinction, Stich seeks to identify features of beliefs that undergird our intuitions about what does and does not count as belief, thereby providing a partial "analysis of our ordinary concept of belief" (1978, 499). Stich emphasizes "two characteristics which beliefs exhibit and subdoxastic states do not: access to consciousness and inferential integration [i.e., a lack of encapsulation]" (1978, 511). Since the states routinely postulated in information-processing cognitive psychology appear to lack these features, they cannot count as genuine beliefs.

One might wonder to what extent a distinction that requires marking with arcane terminology is "entrenched in intuition," as Stich asserts (1978, 499). To some extent, Stich actually seems to pivot between two different distinctions at various points: the distinction between beliefs and non-belief states, and the distinction between beliefs and subdoxastic states proper. Whereas the former

[41] Bermudez (2009) interprets Stich as holding that subdoxastic states lack propositional context altogether. On this strong reading, an alternative argument could be made—one which does not rely on confining the scope of normative constraints to decision-theoretical attitudes: Normativist constraints apply only to propositional attitudes; modular mental states are not propositional; therefore, the constraints do not apply to them. I consider this alternative argument briefly below.

distinction is an everyday one, the latter is a bit more abstruse. As Stich rightly points out, traditionally, it has its home in epistemology, where a distinction is made between non-inferential beliefs and the non-beliefs (sensations, etc.) which lead to their formation without serving as premises for them. Stich is suggesting that the distinction should have a home in cognitive psychology as well. For the states which cognitive psychologists posit, for example, to explain the formation of beliefs manifested in judgments of grammatical intuitions should similarly by regarded as sub-doxastic—non-beliefs that happen to be causally implicated in belief-formation.[42]

But Stich does not give compelling reason to think this. In the present article, Stich exhibits an *a priori* functionalist orientation to the psychological ontology.[43] He clearly thinks that through reflection alone it is possible to arrive at "some property or cluster of properties" that capture the essence of such mental states as belief. However, one can question such an ontology and the semantics on which it is based. If one adopts a Kripkean approach to the semantics of mental vocabulary, then the essence of mental states like belief turns out to be something that is discovered only *a posteriori*. Hence, no special privilege need attach to those predicates which reflect our intuitions about the properties possessed by mental states like belief.

But leaving aside broad issues of ontology and semantics for the moment, Stich's argument fails on its own terms. For the intuitions that beliefs must be accessible to consciousness and inferentially integrated hardly appear robust.

[42] Significantly, however, whereas the subdoxastic states of traditional epistemology lack inferential connection with the beliefs they cause (because they are taken to lack propositional content), Stich regards the subdoxastic states of cognitive psychology as exhibiting inferential relations amongst themselves and with the beliefs they produce.

[43] Cf. (Stich 1983) where Stich argues on the basis of an *a priori* functional analysis of belief to an eliminativism with respect to it.

Personally, I do not share these intuitions in any significant degree,[44] nor—more tellingly—have the abundant cognitive psychologists who have felt little compunction in couching their modular and other information-processing hypotheses involving tacit processes in terms of belief. If access and integration were *a priori* conditions on beliefs, as Stich evidently takes them to be, these theorists—implausibly—would stand convicted of serious conceptual confusion.[45]

Generally, Stich appears over-confident in his intuitions. This is particularly glaring when he attempts to precisify the access-to-consciousness criterion. Wishing to accommodate Freudian unconscious beliefs, he claims that our intuitions can allow for lack of conscious access when this is due, as in the case of psychoanalytic theory, to "a psychological mechanism capable of interfering with the ordinary process leading from belief to assent or to conscious awareness" (1978, 505). But I suspect that few will have a robust intuition that supports drawing a line between belief and non-belief at so seemingly arbitrary point as does Stich, namely, between informational states unconscious due to some blocking mechanism and informational states (like those modular theory posits) unconscious owing to the mind's structural

[44] It is instructive to contrast the—quite strong—intuition that belief does not share the direction of fit of desire to see how weak the intuitions Stich cites really are.

[45] To his credit, Stich at least tries to explain how it is that these psychologists have neglected his intuitions. He writes, "It would be my guess that . . . many of those concerned with cognitive simulation have been so captivated with the promise of inferential accounts of the mechanisms underlying perception and thought that they have failed to note the rather special and largely isolated nature of the inferential processes between beliefs and subdoxastic states. Failure to take seriously the matter of access to consciousness likely has a less creditable explanation. Since the heyday of behaviorism, conscious awareness has had a bad name among many psychologists. And the attitude seems to persist even among those who have come to see behaviorism as a dead end" (1978, 517). But that Stich's intuitions are so easily resisted even when—post-Fodor—the notion of encapsulation has been thoroughly assimilated and behaviorist scruples about consciousness have long receded into the past, belies Stich's explanations.

features.[46] Moreover, Stich plays rather fast and loose with his intuitions. He identifies it as "a further principle embedded in our pre-theoretic notion of belief" that "inference . . . is a relation exclusively among beliefs" (1978, 511). Yet despite holding that inferential relations exist among the states hypothesized in the information-processing models of cognitive psychologists,[47] he nonetheless—inconsistently—insists that they are not beliefs.[48]

It is also worth considering whether (and how) Davidson could appropriate Stich's argument. Presumably, Davidson would not regard Stich's access and integration conditions on belief as constituting a partial analysis of belief in the way that Stich, an *a priori* functionalist, does. For his account of attribution based on the Principle of Charity is as close to an analysis of belief as a normativist like Davidson is able to offer. But could Davidson argue that that account entails either or both of Stich's conditions on belief?

I am not prepared to argue definitively that that account does or does not entail these conditions. But a few observations are in order. First, with respect to the

[46] For all Stich's effort to accommodate Freudian unconscious beliefs, Freudian theory is as replete with completely inaccessible unconscious informational states as merely repressed ones. So, in fact, Stich's access-criterion threatens to consign modules like those on which my argument rests in large part to the sphere of subdoxastic states and, therefore, potentially outside the scope of normativism. Hence, the importance of my addressing Stich's distinction.

[47] Whether such processes should in fact be seen as inferential is an issue which I consider below.

[48] An additional point worth mentioning is the following: Stich takes for granted that the states hypothesized in information-processing models lack inferential integration, and it must be conceded that many have been inclined to see informational encapsulation as an essential feature of modular processes. Carruthers, however, (2006) argues for a picture of modularity that considerably weakens or even banishes altogether this condition on modularity. Accordingly, even if inferential integration were required by belief, this would not clearly disqualify *all* modular states.

There are complications with respect to access-to-consciousness as well. Though such information states as those representing things like grammatical rules would be screened off from consciousness, many others involved in modular hypotheses have more ordinary contents which are (or can be) conscious. Perhaps, though, in the case of the latter Stich might insist that when such contents are conscious it is not the selfsame state that is so. Rather, a conscious (or pre-conscious) belief gives rise to a subdoxastic state with a similar content (or vice versa).

access condition, in (2004c), where Davidson defends other Freudian theses, he also shows himself friendly to the existence of unconscious attitudes (2004c, 185-86). It is not clear, however, whether or not he would countenance regularly inaccessible states as well as merely blocked ones of the sort that Stich addresses. Second, perhaps Davidson's semantic holism vaguely suggests that having a propositional content at all, let alone a belief, requires that it be inferentially integrated with the broader network of such contents. But, more importantly, Davidson's Threshold Principle, which requires than an agent exhibit a preponderance of (ideal) rationality, seems to entail that one's beliefs be highly integrated. Otherwise, one runs the risk of readily falling into inconsistencies or, more generally, routinely drawing theoretical or practical inferences that, though locally rational (i.e., with respect to those attitudes with which they are integrated), considered from the standpoint of one's entire belief set, are less than ideally rational. So quite possibly Davidson would endorse inferential integration as a constraint on belief.[49]

But most important to note is that in the present dialectical context it would be question-begging for Davidson to try to appropriate Stich's argument along the lines just outlined. For at issue is whether Davidson can resist my arguments of Chapters Four and Five against his normativism by (1) insisting that his normativism applies only to beliefs (and other decision-theoretic attitudes) and (2) denying that modular processes involve beliefs. It is clearly question-begging, then, for him to appeal to his normativism to establish that beliefs must be inferentially integrated and that, therefore, modular processes cannot count as beliefs. Hence, I conclude that attempts

[49] Whether this should lead him to deny that modular processes involve belief, however, will depend on just how non-integrated those processes are.

to ward off my arguments against normativism by invoking a belief-subdoxastic distinction fail.[50]

Likewise, this holds true of the alternative counter-argument founded on a strong reading of ‘subdoxastic’ according to which subdoxastic states are interpreted as lacking propositional content altogether.[51] On this reading, the counter-argument runs as follows: Normativist constraints apply only to propositional attitudes; modular mental states are not propositional attitudes; therefore, normativist constraints do not apply to modules. A merit of this version of the argument is that its first premise, unlike that of the original version, is uncontroversially true. This advantage, however, is offset by the difficulty of establishing the stronger second premise. Since considerations drawn from (Stich 1978) fail to establish that modular states are not beliefs, *a fortiori* they fail to establish the argument’s second premise, that they are non-propositional. So this version of the counter-argument is equally unavailing.

Is Normativism Confined to Decision-Theoretic Attitudes?

Suppose for the moment, however, that Stich had succeeded in establishing this much—that modular states are not beliefs (or other decision-theoretic attitudes).

[50] It is important to be clear here: My point is not that the informational states involved in modules should be seen as beliefs. Whether that is so I take to be an open question. Rather, my point is that no compelling *a priori* reason has been given to disallow the possibility that modules might involve beliefs.

[51] It should be noted that, although Bermudez (2009) ascribes the strong reading to Stich, a careful reading of Stich reveals that at no point does he commit himself to subdoxastic states necessarily being non-propositional. Granted, the strong reading is encouraged by the fact that the corresponding distinction drawn in traditional epistemology (between non-inferential beliefs and the non-belief states underpinning them)—on which Stich claims to base his own distinction—is naturally taken as coinciding with a propositional/non-propositional distinction. For the tradition would hold that it is precisely because the non-belief states (like sensations, etc.) lack propositional content that their support of ultimate beliefs is non-inferential. Stich, however, breaks with the traditional distinction in allowing that sub-doxastic states can have inferential relations with beliefs (and among themselves). While this does not necessarily indicate that he takes sub-doxastic states to be propositional (since perhaps inferential relations can subsist among non-propositional items—an issue to which I return below), it does suggest that one cannot mechanically read off every feature of the traditional distinction onto Stich’s own. Certainly, Stich does not explicitly state that subdoxastic states should be construed as non-propositional. Nor does he cite any considerations that can be construed as establishing that they would be.

This would counter my argument against normativism only if the scope of normativism is confined to decision-theoretic attitudes. But is it? Above (see p. 68) I have argued that Davidson apparently intends charity constraints to apply, not just to decision-theoretic attitudes, but to all (familiar) sorts of propositional attitudes. I shall dwell a bit further on this point. As discussed, this reading of the scope of Davidsonian charity is not without its difficulties. In the first place, his most detailed account of attribution (2004e) confines itself to the decision-theoretic attitudes. Nor are attempts to extend his account more broadly unproblematic. Davidson rejects the idea of trying to define other sorts of attitudes in terms of basic, decision-theoretic ones. An alternative approach to extending his account is suggested by George Rey (personal communication). Rey holds that a kind reading of Davidson would see him attributing attitudes in two stages: (1) an initial attribution of decision-theoretic attitudes and (2) a second round, parasitic on the first, in which non-decision-theoretic attitudes are ascribed. But this leaves Davidson with a rather large—unacknowledged—promissory note to make good on, namely, setting forth an account of how the second round of attribution is to procede.[52]

But even leaving aside the textual evidence, the problems presented by trying to restrict the scope of Davidson's account to decision-theoretic attitudes are even greater. For then Davidsonian interpretativism could no longer claim—even in

[52] Rey's proposal raises subtle issues about the relations between Davidson's attributionism, charity, and rationality-evaluability. On the simplest picture, the scope of all three would coincide. But Rey actually proposes his idea of an extended account of attribution as a way Davidson, not merely could account for all types of attitudes, but might even allow for types which need not hew to charity constraints. In that case, attribution would be broader than charity (and, presumably, rationality-evaluability). (I discuss this aspect of Rey's proposal in greater detail in Chapter Four.) Davidson himself seems to regard at least some non-decision-theoretic attitudes as subject to charity (cf. p. 68 above). At any rate, for the time being, I shall assume the correspondence for Davidson of attribution, charity, and rationality-evaluability.

prospect—to amount to a general account of propositional attitudes and propositional content.[53] It would stand in need of supplementation by some other style of account altogether to handle non-decision-theoretic attitudes. But diminishing the ambitions of the theory entails a corresponding diminishment of its interest—and plausibility. For it seems unlikely that there should be two entirely disjoint constitutive bases for propositional content. And to the extent that *some* suitable non-interpretavist basis would need to be sought for the additional propositional attitudes, Davidsonians would face the—perhaps recalcitrant—problem of explaining why that basis is unsuited to serve for the decision-theoretic propositional attitudes as well. So a limitation of the scope of interpretavism renders it unsatisfyingly narrow, implausible, and unstable.[54] Accordingly, all things considered, it appears preferable to take the scope of Davidson's normativism to encompass all attitudes.

Thus, the first premise of the counter-argument based on Stich's belief-subdoxastic distinction should be rejected: It *cannot* be assumed that normativist constraints are intended to apply only to beliefs and other decision-theoretical attitudes. Accordingly, even if modular states are not decision-theoretic, this will not automatically exempt them from having to obey charity constraints.

[53] Davidson is quite explicit, of course, that his attributionism lays no claim to being an account of the mental in general, since, for example, it has nothing to say about mental states besides the attitudes. But on the view presently being canvassed, its significance is still further circumscribed, namely, to a subset of those attitudes.

[54] Most of these problems are avoided on the view that takes the additional attitudes as parasitic on basic ones: The account retains its interest by being an account of all the attitudes (at least in prospect), does not implausibly involve two disjoint accounts of content, nor lay itself open to subversion by a non-normativist alternative. Of course, there would remain the rather tall order of specifying precisely just how the additional attitudes might depend on the basic ones.

An Argument Based on a Distinction from Hornsby

Another, rather similar, counter-argument to my argument against normativism is based on the purported distinction between personal and subpersonal mental states. The counter-argument runs as follows: Normativist constraints apply only to personal states; modular mental states are subpersonal states; therefore, normativist constraints do not apply to modules. To assess this argument, it will be necessary to attain some clarity about the personal-subpersonal distinction itself. The widespread acceptance of some such distinction among many contemporary philosophers of mind warrants dwelling on it at some length.

The Personal-Subpersonal Distinction

Hornsby (1997a) offers a lucid treatment of the distinction. Hornsby observes that the personal-subpersonal terminology originates with Daniel Dennett, who, in turn, finds precedent for the distinction in the work of Ludwig Wittgenstein and Gilbert Ryle (Dennett 1969, 95). This list of names suggests that it is primarily representatives of the hermeneutical approach to the mind who have found it useful to employ personal-subpersonal distinctions. At its most basic, the hermeneutical view posits a dichotomy among the methods of the natural and human sciences. However, with the rise of cognitive science in recent decades, which employs natural-scientific methods, philosophers like Hornsby have had to retreat somewhat from this stark split between the psychological and the natural-scientific. So Hornsby posits instead a distinction *within* the mental realm between folk or commonsense psychology, on the one hand, and scientific psychology, on the other. She insists that, despite the successes of cognitive science, commonsense psychology persists as an autonomous domain, impervious to scientific psychological results and possessing a distinctive ontological and methodological character.[55] Whereas traditionally hermeneutical and naturalistic approaches to the mind are conceived of as competing, on Hornsby's picture, they are viewed as cooperative enterprises. Scientific psychology merely augments commonsense psychology, while preserving it intact. "There is no longer any need," she writes, "to choose between accounts . . ." (1997a, 159). One has the best of both worlds—or so it would seem.

The personal-subpersonal distinction has its place, then, within this project of carving out a domain of commonsense psychology insulated from encroachment by scientific psychology. For Hornsby, Dennett, and others, that distinction serves to mark the Great Divide between folk and scientific psychology: On the one side is the personal level, the domain of familiar, everyday mental states and activities, and on the other, the subpersonal level, the more esoteric province of cognitive science. More needs to be said, however, by way of clarifying that divide, since the terms 'personal' and 'subpersonal' are themselves rather esoteric, and there is no one uncontested way of drawing the distinction.

The Explanatory Distinction

Hornsby, citing the example of Dennett, advocates making the distinction in terms of the sort of explanations mental states permit. Thus, she distinguishes between states which submit to the explanations of scientific psychology (subpersonal) and those which submit to commonsense-psychological explanations (personal). In the case of commonsense-psychological explanations, she clearly accords reason-explanation a prominent place, although she takes the commonsense-, personal level to include sensation and perception, which plainly requires explanation

[55] More particularly, as for Davidson, commonsense psychology has an interpretationist basis that renders it subject to charity constraints.

of some non-rationalizing character: “We constantly treat one another as sentient and rationally motivated: we use commonsense psychology,” she writes (1997a, 158).

As for scientific explanation, presumably she would accord the deductive-nomological pattern of explanation an important place. However, she seems particularly to emphasize the role in cognitive science of what initially appears a version of Robert Cummins-style explanation by functional analysis (Cummins 1983). Thus, she writes, subpersonal accounts “have a place in a story of how it can be that something has the various *capacities* without which nothing could be the sort of intelligible being that a person is” (1997a, 161). (This stress on the explanation of capacities is, of course, characteristic of Cummins-style explanation.). However, as I shall suggest below, it is more likely that she construes cognitive science’s explanatory role in terms of Dennett’s three explanatory stances: intentional, design, and physical (Dennett 1971).

In any case, Hornsby clearly holds that mental (and psychologically relevant physical) states can be partitioned into disjoint classes according to whether they admit of commonsense- or scientific styles of explanation, respectively. She writes, “the accounts of subpersonal Psychology must be addressed to a different set of explananda” than those of personal psychology (1997a, 167). Moreover, on her thoroughly hermeneutical view, the explanatory differences among these classes of states reflect a deep difference of ontological and methodological character: Commonsense- and scientific psychology differ in both their subject matters and methods. Most crucially, the states studied in commonsense-psychology are subject to normativist constraints of the general sort articulated by Davidson.[56]

Daniel Dennett agrees with Hornsby as to essentials. Like Hornsby, in his original employment of the terms ‘personal’ and ‘subpersonal’, Dennett also makes the distinction along explanatory lines. Unlike Hornsby, however, the early Dennett—in a way that now seems dated—limits the subpersonal level to physical states of the human being (1969, esp. 178-79). Hornsby, writing later than Dennett, is in a better position to acknowledge that cognitive psychology trades in both neurophysiological states and fully intentional states. Accordingly, she distinguishes two corresponding kinds of subpersonal psychology, “*neuroscientific*” and “*functional*” (1997a, 163), and contemporary proponents of the personal-subpersonal distinction follow her in this.

The Whole-Part Distinction

But one may wonder what justifies marking the split between commonsense and scientific psychology with the terms ‘personal’ and ‘subpersonal’. Apparently, at least part of what Hornsby has in mind is that only reason-explanations such as figure in commonsense psychology seem to involve an essential reference to a person (1997a, 157-8, 161). In providing a rationalizing explanation of why Columbus sailed the ocean blue one must mention Columbus himself (“He wanted to find a shorter route to the Orient and believed he could do so by sailing west”), whereas in providing scientific explanations of psychological phenomena one can adopt an impersonal causal idiom (“Representations p and q in the visual system produce representation r”).

[56] For Hornsby’s allegiance to a broadly Davidsonian account of meaning and content, with its attendant normativism, see (1997b, 195-220).

Clearly, then, personal-subpersonal distinctions have much to do with subjects of predication, that is, with *which* subjects possess *which* mental and neurophysiological properties. This connection to subjects of predications is explicit in a second personal-subpersonal distinction that Hornsby discusses, if only to reject it. Hornsby observes that the personal-subpersonal distinction is sometimes drawn in terms of a whole-part distinction, personal-level states being those properly predicated of the whole person, subpersonal-level states, of some part of the person. She notes, further, that the whole-part distinction is not equivalent to the explanatory distinction, since, she maintains, "there can be facts about (whole) persons which, because they lack explanations from the commonsense standpoint of commonsense psychology, are not personal-level facts [in the explanatory sense]" (1997a, 162). (Apparently, she means the "capacities and abilities upon which our being commonsense psychological subjects depends," things like the ability to "recognize faces, to catch balls, to do long multiplication . . ." [1997a, 159].) Since, as I have emphasized, her main interest in the personal-subpersonal distinction is to mark out a domain methodologically discontinuous with natural science, she sets aside the whole-part distinction in favor of her and Dennett's explanatory distinction.

However, although she is not explicit, it is fairly clear that she conceives of her distinction as at least in large part coinciding with the whole-part distinction. For at several points she helps herself to formulations that reflect that distinction. Thus, for example, in discussing functional subpersonal accounts of the seemingly simple activity of catching balls, she maintains that the abundant complex tacit calculations such accounts ascribe to a ball-catcher are "carried out inside her; and not *by* her

(1997a, 165). That is, they are activities of some part (or parts) of the ball-catcher, not of the whole person herself. It is instructive, then, to consider how the whole-part distinction divides up mental and neurophysiological states into personal and subpersonal categories.

Proponents of a whole-part distinction have a very definite conception how that distinction carves up such states. On that conception, the personal is supposed to take in conscious mental states, whereas the subpersonal is supposed to consist of (deeply) unconscious states, as well as any neurophysiological states that support mental activity.[57] The affinity in this regard to Stich's belief-subdoxastic distinction is striking. Granted, Stich is more exclusively focused on mental states (and, even more specifically, informational ones). However, with respect to these, he is explicit in placing the boundary between beliefs and subdoxastic states along the conscious/deep unconscious divide.[58] The impression one has is that the belief-subdoxastic and personal-subpersonal distinctions, though notionally quite different—the former is couched in functionalist terms, the latter in terms of subjects of predication (or style of explanation)—give expression to the same root intuition:

[57] The qualifier 'deeply' is required because proponents of the whole-part distinction would assign states that are merely contingently inaccessible to consciousness to the whole person. For example, the repressed states postulated by psychoanalytic theory would be judged personal, since on that theory they are ultimately accessible (though with difficulty). By contrast, the thoroughly inaccessible states of the cognitive unconscious would be classified as subpersonal. For the notion of a deep unconscious see (Searle 1992). On the cognitive unconscious see (Kihlstrom 1999).

[58] As noted (see p. 134 above), he takes access to consciousness as a necessary feature of belief. More precisely, he maintains that our intuitions can allow for lack of conscious access when this is due, as in the case of psychoanalytic theory, to "a psychological mechanism capable of interfering with the ordinary process leading from belief to assent or to conscious awareness" (1978, 505). So it is *deeply* unconscious states subserving belief-formation that he reckons subdoxastic.

It is worth noting, further, that Stich seems to view his other condition on belief, inferential integration, as performing no additional work in determining the extensions of 'belief' and 'sub-doxastic'. Consciousness and inferential integration seem to coincide extensionally for him, at least contingently.

that folk and scientific psychology carve up the mental sphere along the conscious/deep unconscious divide.

But does the whole-part distinction succeed in capturing the intended extensions where Stich's distinction did not? Now perfect sense can be made of a distinction between states and properties of the whole person and of a part of the person. My brain, for example, has the property of being smaller than a breadbasket whereas I myself lack that property. However, there is little reason to suppose that the whole-part distinction corresponds to the categories of conscious and non-conscious states, respectively.

For example, despite Hornsby's claims to the contrary, *can't* one attribute the tacit calculations involved in ball-catching to the whole person, to the ball-catcher as well as, say, some module (or modules) forming part of her mental architecture? To many of us, there will appear nothing incoherent in saying something like, "She (the ball-catcher), albeit unconsciously and involuntarily, performs numerous complex calculations before placing her legs and arms in a position suitable for catching the ball." Why, then, are typical proponents of the personal-subpersonal distinction inclined to say that there is? It cannot be on the basis of some general principle that states properly predicated of mere parts cannot be states of the whole of which they are parts. For anyone will readily admit (quite loudly, in fact) that *they* touched the hot stove, when their hand accidentally comes in contact with it. Though to infer that something *must* belong to a whole because it belongs to its part is to commit a fallacy of composition, it is equally fallacious to infer that what belongs to the part *cannot* belong to the whole. Otherwise, proponents of the personal-subpersonal distinction would be forced to concede that the familiar states (sensations, perceptions, beliefs, desires, etc.) that belong to the part of the person known as the conscious mind are themselves subpersonal, not states of the whole person at all!

So why *do* proponents of the whole-part distinction think that distinction divides mental and neurophysiological states up in the way that they suppose? I suspect that what may be at work here is the following: At times, people are susceptible to an intuition that they (their *self*, the person that they are) begins and ends at the borders of their consciousness; that what lies beyond, the unconscious mental states cognitive scientists attribute to them, are not, in fact, their own. Indeed, someone in the grip of this intuition feels that the entire cognitive unconscious is *no part* of one's self;[59] even more, that the very brain and body to which they are attached is no part of them. Of course, this intuition is a dualist one; but the longevity of that philosophy of mind bespeaks the prevalence and power of this intuition.

This intuition concerns a distinction between the person and the non-person (between my states and the states of what is not even part of me, although associated with my thinking in some way). Moreover, those in the grip of this intuition place the boundary between person and non-person precisely where the adherent of the whole-part distinction does, namely, between conscious and non-conscious states.[60] Perhaps, then, at some level those tempted by the whole-part distinction confuse their personal-*sub*personal distinction with this person-*non*person distinction.

[59] This sense is intensified by the strangeness of the sorts of processes and states involved in accounts of the cognitive unconscious. It reaches its extreme when, as in phenomena like unconscious racism, the contents attributed are ones that we, not just fail to identify with, but would consciously repudiate (cf. Wilson 2002).

[60] Perhaps the distinction is better put as one between self and non-self. This is because what is no part of me may be part (or parcel) of some other person. Indeed, one can be drawn to a view that sees the cognitive unconscious as part of some *other* person than oneself. Cf. the related discussion in (Marks 1981).

They may also be misled by a specious argument along the lines of the following:

(1) The cognitive unconscious (and other such parts of minds) are essentially minds in miniature.

(2) Token mental states cannot be shared by different minds.

(3) Therefore, a person cannot share in the token mental states of the cognitive unconscious.

The first premise may or may not be true; the issues involved are complex and best addressed in another context. The second premise, of course, is one that many in the history of philosophy have found compelling, perhaps even a conceptual truth. But even supposing both premises true, the argument is, of course, invalid. It can be *rendered* valid by supplying the additional premise that persons are minds, but that is not something one can take for granted. Indeed, on the usual view, a person is neither identical to, nor (at least typically) constituted by, a mind. Rather, they are in some sense composites of both mind and body.[61] So the argument under consideration fails to justify the conclusion that deeply unconscious states cannot be states of the whole person, as the proponent of the whole-part distinction believes. In fact, if the argument were sound, it would boomerang against the proponent of the whole-part distinction. Applied to the conscious part of the mind it would yield the conclusion that persons cannot share in the token mental states of the conscious mind either! Again, there seems little reason to think that a whole-part distinction coincides with a distinction between conscious and unconscious states.

[61] The intuition that one's self consists in one's consciousness may lull some, however, into accepting that persons are minds. That intuition, though question-begging in the present context, may further explain why some are taken in by the present argument.

Accordingly, it is unavailing to the normativist who hopes to evade the force of my argument against normativism by confining the sphere of application of charity principles to personal states. For the whole-part distinction fails to exclude the unconscious subsystems on which my argument relies from the putatively charity-governed personal sphere. But perhaps the explanatory distinction will fare better in making the desired cut.

Assessing the Explanatory Distinction

There are grounds, however, for doubting the very tenability of the explanatory distinction. As noted, Hornsby holds that her explanatory distinction bifurcates mental states into those that allow of only scientific explanation (subpersonal) and those that allow of only commonsense-explanation, especially reason-explanation (personal). Whereas the former possess the status of scientific posits, the latter demand an interpretationist account, one which subjects them to charity constraints.

But there is cause to harbor some doubt about Hornsby's bifurcation. Perhaps one and the same state can admit of both rationalizing *and* scientific explanation, say, deductive-nomological explanation. As I argue below (p. 158, and n. 68), systems like those posited by cognitive scientists to explain human reasoning seem to involve processes that permit both ordinary reason-explanations and purely causal, information-processing ones. So there is some reason to question Hornsby's bifurcation of states by style of explanation. But perhaps if pressed, Hornsby could simply recast her explanatory distinction in terms of those that admit of scientific styles of explanation, on the one hand, and those that admit of *only* reason-explanation, on the other. This would allow her to maintain a split between those states that are scientifically-tractable and those that are not, despite the consideration that I have just raised, and this might suffice for her purpose of marking out an autonomous domain of commonsense psychology.[62] However, there is ample reason to doubt whether commonsense psychology is secure against the inroads of scientific patterns of explanation.

In the first place, it may be that folk-psychological explanation itself can be cast in deductive-nomological terms. Many think that folk psychology is a shared (possibly innate) tacit theory which people use to explain one another's behavior along the lines of the scientific model (see, e.g., Gopnik and Melzoff 1997). The reason-explanations that people cite explicitly may simply be elliptical expressions of unarticulated underlying explanations fitting the D-N model. Such a view, if true, would completely undermine Hornsby's insistence on the explanatory distinctiveness of folk psychology.

Of course, the theory-theory view of folk psychology on which such a conclusion rests is controversial. Moreover, in fairness to Hornsby, it must be granted that, *prima facie*, much in the sphere of folk psychology may look out of reach of scientific explanation. Scientific psychological treatments tend to be directed at the workings of subsystems of the mind or person, leaving the field largely clear for commonsensical explanations of the large-scale *behavior* of the person.

But this is not to concede the *impossibility* of scientific-psychological explanation at that level. Behavior at that level would almost certainly be the

[62] Though she would have to concede that at least many cognitive-scientific processes permit reason-explanation and, therefore, are 'rationality-evaluable' (i.e., permitting of evaluation by rational standards), she could, nonetheless, consistently maintain that they are immune to charity requirements. So the modified personal-subpersonal distinction could still serve to bracket off irrational subsystems from the scope of charity.

complex resultant of the interactions of numerous psychological subsystems. It may very well be, therefore, that few, if any, law-like generalizations can be formulated that apply directly to that large-scale level which, in turn, would permit employment in standard deductive-nomological explanations. But given laws applying to the subsystems and a model of how they interact, it might be possible in principle, however much difficult in practice, to provide scientific explanations of an individual's large-scale behavior. In that case, rationalizing or other commonsensical explanations would at best amount to practically indispensable stopgaps, to be employed pending attainment to fuller, more adequate scientific explanations.

Moreover, when one turns from the human being's overt behavior to their familiar mental states (their beliefs, desires, perceptions, etc.), the prospects for scientific explanation appear much improved. As I intimated above (p. 144), Hornsby seems beholden to Dennett's picture of cognitive science. But that picture, I think, misrepresents its true character. On that well-known view, human beings' behavior in the broadest sense permits explanation at three independent levels, the intentional-, design-, and physical-levels (Dennett 1971). Commonsense psychology, with its vocabulary of familiar states and its rationalizing pattern of explanation, operates at the intentional-level, whereas cognitive science stakes out the lower design- and physical-levels. Perhaps Hornsby would see the two sorts of subpersonal psychology she mentions, functional and neurophysiological, as corresponding to these two levels, respectively.

Clearly, this picture lends itself wonderfully to the view that commonsense psychology is an autonomous enterprise, insulated from scientific psychology. For on this picture, each level addresses its own proprietary set of explananda, with the methods and concepts peculiar to that level. Cognitive science, thus banished to the design- and physical-levels, leaves folk psychology, occupying the intentional-level, intact.

This tidy picture suggests that the states which cognitive science trades in are taxonomically distinct from those of commonsense psychology. Indeed, Hornsby is quite explicit in contrasting "belief and desire," for example, with the "states of a subpersonal Psychology," however great the "informational, or 'cognitive', sophistication" of those latter states may be (1997a, 164-65). But the fact that psychological theorists with no particular philosophical axe to grind so readily apply the familiar categories of belief, desire, etc., to the informational and desiderative states that figure in their modular (and other) accounts suggests that the burden of proof is on proponents of Hornsby's distinction to show that the unconscious states these accounts treat of are not of familiar sorts.[63] On its face, a personal-subpersonal distinction appears a needless reduplication of types of mental state: for every familiar sort of mental state, proponents of that distinction are forced, uneconomically, to posit an analogue on the other side of the conscious-unconscious divide. But why should not instead one and the same familiar sort of mental state be capable of enjoying a dual life on each side of that divide?

Moreover, Hornsby's tidy picture on which commonsense- and scientific psychology operate at distinct explanatory levels fails to do justice to the way cognitive science actually works. *Pace* Hornsby and Dennett, the explanatory models

[63] Hornsby clearly allows that such states include propositional attitudes, indeed, ones "whose contents are of the right sort to be contents of states of ourselves" (1997a, 166).

which cognitive scientists typically formulate do *not* treat of some set of explananda altogether isolated from those of commonsense psychology. For it must be allowed that cognitive psychology seeks to explain familiar kinds of states like conscious sensations, perceptions, beliefs, and desires (such as some would characterize as personal), even if in doing so it appeals to an apparatus of, in some respects markedly different, deeply unconscious states and processes (which some might be tempted to label 'subpersonal'). The abstruse posits of cognitive scientists are introduced to causally explain conscious perceptions, beliefs, etc. Thus, for example, a module that parses sentences along Chomskyean lines, on the basis of quite ordinary perceptual information as input delivers a quite ordinary grammatical judgment as output (Chomsky 1980). Even if the intervening steps be thought to involve reference to unfamiliar sorts of states, they must be thought of as enmeshed in a complex causal interaction with familiar sorts. Hornsby and Dennett's picture falsifies both scientific and commonsense-psychology by neglecting their close relationship.

Moreover, the inescapable overlap among commonsensical and scientific psychology—the fact that conscious perceptions, beliefs, etc., need to figure in both forms of psychology—threatens the very consistency of Hornsby's personal-subpersonal distinction. For Hornsby's distinction is supposed to mark an ontological and methodological divide; and even if there is no problem imagining a type of state (belief, say) figuring in both commonsensical and scientific explanations, that ordinary types of states would need to figure in both commonsense and scientific psychology would (on the hermeneuticist's assumptions) require them—*per impossibile*—to embody two distinct constitutive and methodological natures simultaneously! In sum, a personal-subpersonal distinction such as Hornsby wants to draw appears untenable. Accordingly, it fails to afford the normativist a coherent basis on which the scope of charity can be restricted to a subset of propositional attitudes and, therefore, can be of no real use to the normativist who wishes to evade the force of my anti-normativist argument.

Are Modular Processes Non-Inferential?

Another attempt to counter my argument against normativism might seek to deny the relevance of modular processes to charity principles by denying that standards of rationality even apply to such processes, that is, by denying that they are 'rationality-evaluable'. In particular, I mean to consider a possible argument for this conclusion based on denying that modular processes are *inferential*: If the transitions between states of modules are non-inferential, then they would seem not to permit reason-explanation and, therefore, fall out of the sphere of evaluation by rational norms altogether.

The notion that modular processes are not inferential might be suggested from certain considerations mentioned by Zenon Pylyshyn (2003). Whereas Fodor's usage of 'inference' (e.g., in Fodor 1983) is quite liberal, Pylyshyn does not think it appropriately employed "to refer to processes that are systematically restricted as to what type of input they may take and the type of principles that they follow" (2003, 38n8). Thus, he wishes to distinguish inference from "other sorts of causal regularities" on the basis of (at least) two criteria. Inferential processes must apply to representations from all domains and they must embody distinctively *logical* principles. Although Pylyshyn is particularly concerned to deny that the processes involved in early vision are inferential (since they take only visual inputs and embody non-logical processes), his line of reasoning might be thought to impact other putatively modular processes as well.[64] So it merits some consideration.

Many sorts of modules postulated by theorists besides early vision follow principles which appear non-logical. Thus, desire-forming modules produce desires on the basis of various inputted beliefs, etc., yield desires as outputs, where there can hardly be said to be logical principles which license the transition from those beliefs, etc., to the desire.[65] Moreover, domain-specificity, if not exactly a necessary condition for modularity, is characteristic of most modules posited by contemporary theorists. In fact, at least for proponents of massive modularity, the bulk of deductive and statistical inferences are supposed to be carried out by domain-specific modules.

[64] One can envision another route to the specific conclusion that early vision is not inferential. In logic, 'inference' is, of course, a synonym of 'argument', where this is typically understood as a structure consisting of propositions. So if one accepts Fodor's line that early-visual contents are not propositional, then, guided by the logical usage, one may be inclined to conclude that early-visual processes are non-inferential. With respect to putatively modular processes more broadly, the intuition (discussed by Stich but, ultimately, rejected—see p. 137 above) that inference is a relation only among beliefs will lead one to conclude that modular processes are non-inferential if they do not involve beliefs (say, because they are subdoxastic). I have argued there is no compelling reason to accept the notion that modular processes are subdoxastic. But the intuition would still entail that non-propositional, Fodorian early vision is not inferential since it cannot involve belief, which is uncontroversially propositional. Moreover, the intuition would yield the same conclusion even with respect to modules whose processes involve any sort of propositional attitude besides belief (e.g., desire-forming modules, for which see Carruthers [2006]). In this thesis, I do not assess whether inference subsists only among beliefs or propositional attitudes, since, as will emerge, I think the rationality-evaluability of modules can be defended even if they are non-inferential.

[65] In the first instance, one might conclude this because the output of such modules is a desire and, one might hold, logical principles only apply to beliefs. However, if one takes a less narrow view of the 'logical principles' which, according to Pylyshyn, must figure in inferences, then perhaps they can involve non-informational states like desires and intentions. Pylyshyn's focus is theoretical, but to allow for practical reasoning, Pylyshyn, it seems, should grant that processes can count as inferential in virtue of embodying any sort of principle of rationality, not just those that are logical in character. In that case, the conclusion that the desire-forming modules do not involve inference is much less straightforward. For, presumably, Pylyshyn does not want to require that the principles underlying inference be normatively correct ones. That is, unless he is a normativist, he will not wish *a priori* to preclude the possibility of regular patterns of *bad* inference. So on what basis is one to say that the desire-forming modules do not operate according to logical or rational principles as opposed to normatively flawed ones? I return to this question below.

By Pylyshyn's criteria, however, these could not count as genuine 'inferences'.[66] Is there perhaps something amiss with Pylyshyn's criteria?

It should be noted first that Pylshyn does not really provide any argument for his criteria of inference. But there is a certain—at least *prima facie*—force to his assertion that inferential processes must involve distinctively logical principles: Isn't it logic, after all, that treats of inference?[67] But it is much less clear why one should insist that processes be unrestricted in their inputs if they are to count as inferential. Imagine a module that takes inputs only from some specific domain like folk biology and then issues in beliefs of the form '*a* is *M'* when the inputs include beliefs of the forms 'All *S* are *M*' and '*a* is *S*'. It seems quite natural to label this process an inference, even though if—say, by some alteration of its connections with other modules—it were fed inputs from some other domain, it would operate quite differently or even yield no output at all. Indeed, it seems quite possible to provide a reason-explanation of why the bearer of these states comes to believe *a* is *M*, namely, because she believes that all *S* are *M* and that a is *S*.[68] It seems arbitrary to deny that there is an inference here.[69]

[66] In fairness to Pylyshyn, it should be noted that his attention is confined to vision. It is not clear whether he contemplates the possibility that 'cognition' (in contrast to input systems) might be modular as well. But, in any case, it is appropriate to ask what the conditions he places on inference imply with respect to modules involving central processes.

[67] There may be some difficulty in clarifying just what is and is not a logical principle, as opposed to a bad logical principle, but this need not be an insoluble problem. (Deductive logical principles at least, I suppose, could be characterized as ones involving only topic-neutral vocabulary.)

[68] It is this seeming possibility of applying such reason-explanations to modules which suggests that Hornsby (cf. p. 132 above) cannot define personal-level mental states simply as those that permit reason-explanation. Rather, it appears, she needs to define them as those that admit *only* of reason-explanation (and not also of scientific explanation).

[69] Perhaps Pylyshyn is guilty of a confusion. If processes' inputs are domain-specific, that might be thought to preclude the principles on which they operate from counting as logical, since logical principles are supposed to be domain-general. But such principles are domain-general in the special sense of involving only topic-neutral vocabulary. It is not clear that input-restrictions interfere with their topic-neutrality.

Ultimately, I do not wish to take a definitive stand as to whether (or which) modular processes are inferential. Given my general scientific-essentialist orientation, I am inclined to think constraints on what counts as inference will need to be determined *a posteriori*, not *a priori*, as Pylyshyn attempts to determine them. But suppose Pylyshyn's criteria of inference are correct. Even though this would rule out most modular processes from counting as inferential, I do not think this would have the implication, fatal to my argument, that modules are not rationality-evaluable and, therefore, beyond the scope of charity constraints. This is so for at least two reasons. First, rational standards are statal as well as procedural. So even if modular processes themselves are exempt from rational norms, the states which they output (and even the intervening states bound up in their internal operations) would appear subject to such norms as the requirement of logical consistency. Moreover, the processes themselves, even if non-inferential, will be subject to *epistemological* norms governing belief-formation, which apply regardless of whether the relevant processes are inferential.[70] Hence, charity principles can still get a grip even with respect to modules that involve non-inferential processing. So I conclude that the potential counter-argument to my argument against normativism that attempts to exempt modules from the scope of charity by consigning them to a special realm of arationality fails.

In fact, as my discussion of Stich, Hornsby, and Pylyshyn has shown, attempts to remove modular processes from the scope of charity quite generally appear unpromising. That is, to the extent one is inclined to view propositional attitudes as

[70] Some, e.g., Stephen Stich (1990), interpret rationality broadly to include epistemological norms. But whether one places the latter within or alongside the set of rational norms, normativists like Davidson intend charity principles to encompass epistemological requirements.

subject to charity constraints, there will be little principled basis for denying that modular processes can be so constrained as well. Accordingly, the tenability of those very constraints is rendered dubious by the scientific possibility of modules that do not adhere to them. In Chapter Four, I proceed to describe one such module in detail.

Chapter Four: An Argument From Basic Freudian Wish-Fulfilment

Introduction

With characterizations of charity and modularity in place, I can begin to develop my argument against normativism by describing a module which embodies some of the key processes which Freud ascribes to the part of the personality which he labels the 'id',[1] namely, basic forms of what Freud refers to as 'wish-fulfilment'. I will argue that this module conflicts with the principle of charity, at least the influential version of it proposed by Davidson. Moreover, given the evident coherence of this module, I shall suggest, there is reason to reject normativism's insistence on the principle of charity as a condition of agency. Finally, I shall respond to objections that might be raised specifically against the employment my argument makes of Freudian theory and phenomena.

Freud's Id

Freud introduced the word 'id' (or, rather, its German equivalent, '*das Es*') relatively late in his career (see, esp., Freud 1923). Prior to his formulation of a tripartite division of the personality into id, ego, and superego, he worked with a picture of the personality as consisting of consciousness and the Unconscious, or as Freud sometimes refers to these parts of the personality, the systems *Cs*. and *Ucs*. (see, e.g., Freud 1915).[2] "The nucleus of the *Ucs*.," Freud writes, ". . . consists of

[1] Freud uses 'primary processing' as a general term to refer to the sorts of mental processes involved in the id.

[2] Freud's earlier view of mental architecture is sometimes referred to as his 'topographical' model of the mind, whereas the later, three-part view, as his 'structural' model. With respect to the former, however, it should be noted that he sometimes refers to the system *Cs*. as the 'preconscious' (*Ps*.), since what is essential to it is not the actual phenomenal consciousness of its contents but, rather,

wishful impulses" (1915, 186). It is to the system *Ucs*., further, that unacceptable mental contents are consigned through the repressing agency of the system *Cs*. But since the processes involved in repression (as opposed to what might be termed deliberate 'suppression') are themselves unconscious, it gradually dawns on Freud that it only confuses the issue to label the system to which they belong '*Cs*.' The classification of parts of the personality that he has really intended to make all along is not fundamentally one based on access to consciousness. So 'system *Cs*.' yields to 'ego', and 'system *Ucs.*' to 'id'.[3] But though the terminology changes, the referents (by and large) do not. Accordingly, in characterizing the id, I indifferently apply Freud's pronouncements about the system *Ucs.* to the id.

The id, the ontogenetically earliest system, is the seat of various biological instincts or needs (hunger, for example).[4] Its function is to rid the organism of psychic energy or tension produced by internal and external stimulation. At first, the id exists as a mere sensory-motor mechanism which releases tension (and wards off further stimulation) through reflex actions. But this mechanism ultimately proves insufficient to quell such sources of tension as hunger. So a psychological development ensues. *Primary process* comes into being. 'Primary processing'—it is primary in point of time—is Freud's blanket term for the distinctive mode of functioning of the id.[5] Within primary process, memory-images of instinctual objects

their—relatively—ready capacity for consciousness (1915, 173). The distinction between *Cs*. and *Ucs*., then, should be understood as largely one of access.

[3] On Freud's considered view, then, ego states correlate only imperfectly with (access-)conscious ones and id states with (access-)unconscious ones.

[4] The brief sketch in this paragraph of some of the central features of the id leans heavily on Hall (1954, 15-21).

[5] At times, however (cf. 1915, 187), Freud uses 'primary process' specifically with reference to two processes, displacement and condensation, which he regards as "distinguishing marks" of primary process (186). I discuss these below.

such as food are produced in order to satisfy the wishes caused by instinctual demands. They can do so because the id does not distinguish between such memory-images and genuine perception. It is this representation of "the wish fulfilled as a hallucinatory experience" which Freud terms *wish-fulfilment* (1916-17, 129) The id, operating according to the *pleasure principle*, strives "towards gaining pleasure," with an "entire disregard of reality-testing" (1911, 219, 225). The unconscious processes of the id "equate reality of thought with external actuality, and wishes with their fulfilment—with the event . . ." (1911, 225). As such primary process itself proves inadequate to meet the individual's instinctual demands, *secondary process*, the logical, reality-oriented patterns of thought belonging to the ego, develops. But primary process is by no means supplanted altogether. In circumstances where the ego is unable satisfactorily to minister to one's wishes, primary process revives. Such is the case, for example, during sleep, when primary process produces hallucinatory images in the form of dreams.[6] Moreover, Freud sees wish-fulfilment as implicated in other processes as well, such as neurotic symptoms.

Aside from the id's wish-fulfilling character, its blithe disregard for reality, Freud assigns several other distinctive features to the id and its processes. As Freud writes, the "latent processes" of the id have "characteristics and peculiarities which seem alien to us, or even incredible, and which run directly counter to the attributes of consciousness with which we are familiar" (1915, 170). He maintains, first, that within the id there is "*exemption from mutual contradiction*" (1915, 186). But despite initial appearances, his point does not seem to be that the id freely forms

[6] More strictly, primary process—the id—produces dreams as conscious byproducts of its unconscious activity (see p. 169, n. 19 below).

contradictory beliefs. Rather, as emerges from the following, his point seems to be that the id readily allows conflicting wishes to exist side by side without either being gratified in preference to the other: "When two wishful impulses whose aims must appear to us incompatible become simultaneously active, the two impulses do not diminish each other or cancel each other out, but combine to form an intermediate aim, a compromise" (1915, 186).[7] From the standpoint of charity, of course, this tolerance of conflicting wishes is a much less drastic feature than the tolerance of contradictory beliefs would be.

Moreover, it appears that beliefs within the id are unhedged, that within the id there is "no negation, no doubt, no degrees of certainty" (1915, 186).[8] Again, Freud maintains that the id's processes are timeless: "they are not ordered temporally, are not altered by the passage of time; they have no reference to time at all" (1915, 187). Presumably, he means that the id's contents are untensed, that the id lacks temporal concepts altogether.[9] Additionally, the id is marked by what Freud labels 'mobility of cathexes', by which at bottom he seems to mean that the strength of wishes within the id is variable, with one wish capable either of surrendering its strength to another ('displacement') or appropriating the strength attaching to several others ('condensation') (1915, 186-87).

Doubtless, much exegetical work would be needed to formulate and defend a definitive interpretation of Freud's pronouncements on the characteristics of the id.

[7] If within the process of wish-fulfilment a wish that *p* invariably gave rise to a belief that *p*—along lines which I expound below—then the existence of conflicting wishes in the id would compel Freud to acknowledge contradictory beliefs there as well. But Freud's remark that in such cases of conflict a compromise is sought averts this commitment.

[8] Perhaps Freud also means to assert that the id lacks the concept of logical negation, although it is not entirely clear in the present passage.

[9] Although perhaps he means also to assert that, once formed, mental states persist unaltered in the id?

But I shall spare myself that effort, since I shall be selective with respect to the features of the id that I shall assign the charity-violating module upon which my argument against normativism turns. It is primarily the id's wish-fulfilling character that is reflected in the module that I shall describe. The cogency of my argument in no way depends upon my fidelity to Freudian ideas in all respects. In fact, inasmuch as some have found Freud's full conception of the id dubiously coherent, my argument can only gain in force if the relevant module draws modestly from the set of characteristics Freud attaches to the id. Any doubts an impartial judge might harbor about the coherence of Freud's id, I am confident, will lapse with respect to the module I describe, which, again, seeks to embody Freud's central ideas about wish-fulfilment.[10]

Basic Wish-fulfilment

Because the case for the conflict between Charity and a wish-fulfilling module can be made without bringing in the less basic sorts of wish-fulfilment, I shall initially only treat the two most basic sorts mentioned above: infantile hallucination and dreaming. And, with respect to the latter, I shall only mention the dreams of young children, which in Freud's view, are straightforward, undisguised wish-fulfilments (1916-17, 126-35). Only subsequently, in discussing less basic forms of wish-fulfilment, will I treat the "dreamwork," that is, the various processes by which according to Freudian theory the wishes instigating dreams in adults (the dream's

[10] Linda Brakel (2002) offers a defense of the coherence of the id even construed as possessing most of the striking properties Freud assigns it. She argues for its coherence by demonstrating its consistency with a certain atomistic account of content, namely, Ruth Millikan's proper-function naturalism (cf. Millikan, 1993). In the present context, I take no stand with respect to the success of her attempted defense. (A complication is that her enumeration of the features of the id—and her interpretation of those features—differs somewhat from that which I have offered.)

"latent content") are rendered unrecognizable in the dream-imagery which expresses them ("manifest content").

Sebastian Gardner (1993) presents a thoughtful reconstruction of Freud's theory of the basic forms of wish-fulfilment.[11] He extracts something like the following structure for such wish-fulfilment: An unsatisfied biological need, (1), yields a wish, (2), which, in turn, produces a wish-fulfilling representation, (3), causing a feeling of satisfaction,[12] (4), that, finally, leads to the termination (or quiescence) of the wish, (5) (1993, 124-25). Moreover, Gardner lays stress on several features of his reconstruction. First, the wish-fulfilling representation, (3), as an hallucinatory experience, is sensory in character; and the feeling of satisfaction, (4), is an experience of sense-pleasure. Thus, neither introduces an element of judgment or belief into wish-fulfilment. Although the wish-fulfilling representation possesses a content, it is of a pre-propositional character (1993, 122).

[11] See esp. 120-26. Gardner credits James Hopkins for the conception of wish-fulfilment he develops (124n18).

[12] With respect to the words 'satisfaction' and 'fulfilment', some terminological clarification is in order. First, it is worth noting that though the English 'wish-fulfilment' is used to translate Freud's '*Wunschbefriedigung*', the second element of that compound is usually rendered as 'satisfaction' in English (although sometimes as 'fulfilment'). Second, in common parlance, 'satisfaction' and 'fulfilment' seem virtually synonymous. Third, I think it is safe to say that in the phenomenon of wish-fulfilment wishes are neither satisfied nor fulfilled in any *ordinary* sense.
Acknowledging this, Richard Wollheim introduces a terminological distinction between what he calls 'satisfaction' and 'gratification' (Wollheim 1979, 47). He defines the terms roughly as follows:

> my desire that *p* is *satisfied* iff *p*
> my desire that *p* is *gratified* iff it is for me as if *p*

With the former, Wollheim appears to want to capture ordinary usage of 'satisfaction', but with the latter, the "kind of pseudo-satisfaction," as he puts it, that Freudian wish-fulfilment represents (for a related notion, which Peter Carruthers calls 'quasi-satisfaction', see Carruthers [2006, 297]).
I largely leave to one side for the moment what Wollheim thinks it takes for it to be "for me as if *p*". But at least part of what he has in mind is what I, following Gardner, call 'the feeling of satisfaction' in my reconstruction of wish-fulfilment. This is just the familiar sort of pleasure that often accompanies the actual satisfaction of desires (though not universally, of course).

Such wish-fulfilment is usefully contrasted with what Davidson in (2004a) describes as “wishful thinking” on “a minimal account.” On such an account, wishful thinking is “a case of believing something because one wishes it were true” (2004a, 205). For example, it seems one can imagine a phenomenon of a structure similar to that which Gardner sketches in which wishes immediately give rise to beliefs of identical propositional content (call this phenomenon WT.)[13] Such would qualify as wishful thinking on the minimal account but not as wish-fulfilment, which seems to essentially involve a sensory element. In fact, Gardner emphasizes that basic forms of wish-fulfilment need not be supposed to involve propositional belief at all. He observes, “it is not *a priori* that belief is a condition for the cessation” of the wishes involved; “and there is no independent reason . . . for introducing belief into the process,” since “there are no further effects to be explained that would provide evidence for the presence” of a belief (1993, 125).

Nor, in Gardner’s view, should the ‘wishes’ involved in wish-fulfilment be understood as ordinary wishes or, indeed, any other familiar desiderative propositional attitude: “the psychoanalytic concept of wish is not the same as, however closely it may be related to, the concept expressed by ordinary use of the term” (1993, 126). Indeed, he denies that psychoanalytic wishes are propositional attitudes at all.[14] The upshot, then, is that on Gardner’s reconstruction at least, basic forms of “wish-fulfilment” involve “only pre-propositional content” (1993, 122).

Now in certain respects, Gardner’s account of basic wish-fulfilment is not inimical to my aim of demonstrating a conflict between Freudian wish-fulfilment and

[13] I shall return to this point below.
[14] I shall examine Gardner’s reasons for these views about psychoanalytic wishes shortly.

Davidsonian Charity. First, Gardner appears to conceive of wish-fulfilment as a sort of competence. The impression he gives is of wish-fulfilment occurring as a *ceteris paribus* regularity in circumstances where the straightforward meeting of biological needs is impossible (as for the infant or dreamer). Moreover, on his conception, the relevant competence is not a rational one in that the phenomenon of wish-fulfilment involves non-rationalizing transitions among mental states. But it is not an *ir*rational competence, for the states which it involves are not even the sort which are aptly assessed with respect to rationality, namely, propositional attitudes.[15] Rational norms apply only to sets of propositional attitudes. In fact, for the most part, rational and epistemic norms apply only to sets of mental states which include belief or some type of cognitive attitude possessing a direction of fit (unlike emotion) and, specifically, that of belief.[16]

Comparison with the sort of wishful thinking I discuss above (WT), in which *beliefs* are directly produced by wishes, highlights this point. If the wishes involved are understood to be propositional attitudes, it is immediately apparent, I think, that the structure is an irrational one.[17] (And even where they are not so understood, the structure at least clearly contravenes epistemic norms.) But, when belief drops out, replaced by non-propositional sensory experience, as in Gardner's wish-fulfilment, no

[15] Recall that charity on my reading includes epistemic as well as properly rational norms. Now one might think at first sight that Gardner's wish-fulfilment involved the violation of some sort of epistemic norm. But though the generation of hallucinatory imagery based on wish-like states may seem like poor epistemic design, it would not seem to violate any epistemic norms, which solely govern belief-formation and its warrant.

[16] An apparent exception is the principle that preferences should observe transitivity.

[17] In the absence of anything like a theory which codifies rational norms, such appeals to intuition seem unavoidable.

norms are violated.[18] So wish-fulfilment as Gardner characterizes it ultimately fails to advance my argument against Davidson's interpretavism: If wish-fulfilment is to conflict with Charity, it must involve propositional attitudes (including cognitive ones) in some fashion.

Accordingly, on my somewhat revised account, we are to envision wish-fulfilment as possessing the following structure: An unsatisfied biological need, (1), yields a wish that the need be satisfied, (2), which, in turn, produces an appropriate wish-fulfilling hallucination, (3), resulting in a belief that the wish is satisfied, (4), which causes a feeling of satisfaction, (5), that, finally, leads to the termination (or quiescence) of the wish, (6). I shall say something shortly by way of defending this alternative account (and demonstrating its conflict with charity requirements). But for the moment, then, we are to envision a module that embodies basic processes of wish-fulfilment along the lines sketched. Following Freud's account of the id, the module should be thought of as existing and operating prior to the individual's development of conscious thought altogether. Even with the first blush of conscious thought, however, it can be envisioned as operating alongside conscious thought (and action), especially, in very literal wish-fulfilling dreams generated during sleep when one is unable to minister to one's wishes through conscious action).[19]

[18] No norms are violated regardless of whether the wishes involved are propositional or not: it is the absence of the cognitive attitude that tells.

[19] Note that the dreams themselves should be thought of as falling outside the province of the wish-fulfilling module itself. For though we often speak of someone sleeping as 'unconscious', dreams actually would seem to belong to access-consciousness, whereas I follow Freud in conceiving of this id-module as lacking such consciousness. Hence, the dreams, I think, should be thought of as conscious byproducts of the unconscious hallucinatory experiences which figure in the wish-fulfilling processes proper. I shall address how it is that those experiences are able to become conscious (or, rather, give rise to conscious by-products) in discussing less basic forms of wish-fulfilment in the next chapter.

But in terms of the typology presented earlier (see p. 129 above), what sort of module would this be? Well, the first thing to note is that it is not an informational module like Chomskyean linguistic competence (i.e., a body of thematically related information drawn upon in the performance of specific tasks) but, rather, a sort of processing module—a system charged with performing a specific mental function, namely, gratifying wishes engendered by biological needs.[20] Moreover, it does not seem to be especially Fodorian, especially in view of the role that fully conceptualized content plays within it.[21] Rather, it is a module in the looser sense that Carruthers (2006) employs in arguing his thesis of massive modularity (p. 128 above). The module, further, trades on an additional feature of Carruthers' account of modules, namely, that they can be complex, built up out of more fundamental modular building-blocks. For the id-module I have described is naturally seen as encorporating a wish-generating sub-module (producing wishes in response to information concerning bodily states such as an empty stomach), an image-generating sub-module (giving rise to hallucinations), a belief-generating module, and an affective module (producing a pleasurable response to the wish-fulfilling belief).

[20] As noted earlier, I avoid consideration of the philosophical issue whether such modules are best seen in computational or non-computational terms, as well as the empirical issue of whether one should expect them to be realized by specific neurological systems. It is worth mentioning, however, that interesting work has been done in modeling central aspects of Freudian theory along computationalist lines. See, e.g., Boden (1987) and Wegman (1985).

[21] Perhaps, though, it can be envisioned as possessing at least one of the other defining properties of Fodor modules. At least in its initial form, the module appears domain-specific, since its inputs are limited to wishes derived from the domain of biological need. But it is neither encapsulated nor inaccessible. Although this is concealed prior to the development of consciousness, once consciousness comes on-line the module's porous boundaries becomes apparent. E.g., the beliefs generated by the module during dreams should be seen as passed along to consciousness. I return to the issue of the degree and nature of encapsulation and inaccessibility in wish-fulfilment when considering its less basic forms below. It is worth noting, however, that Freud himself emphasizes that there is considerable "communication between the two systems," *Ucs*. and *Pcs*. (the preconscious). He writes, the *Ucs*. "is accessible to the impressions of life"; it "constantly influences the *Pcs*., and is even, for its part, subjected to influences from the *Pcs*" (1915, 190).

This conceivability (indeed, the scientific possibility) of such a wish-fulfilling module, I submit, should call into question the legitimacy of the normativist constraints that Davidson places on the possession of propositional content. For in the first place, such a module, in which we find a regular tendency for a propositional wish that *p* to produce a sensory wish-fulfilling experience and a propositional belief that *p,* embodies what to all appearances is an irrational competence and, therefore, conflicts with Davidson's Competence Principle.[22] Moreover, such a regular divergence within a single mental compartment from standards of ideal rational rationality fairly clearly conflicts with Davidson's Compartment Principle as well, since that divergence is owing to the module's internal operations and not the external influence of some other mental compartment.[23] Again, corresponding to Freud's view that the processes of the id precede the development of the ego, I suggest that one can envision the wish-fulfilling module as developing prior to those systems which embody in the individual logical, reality-oriented patterns of thought. Prior to the development of those latter systems, the wish-fulfilling module would largely exhaust the individual's capacities for propositional thought and reasoning. In such a mind, irrational processes would clearly predominate and, therefore, the mind itself would appear to violate Davidson's Threshold Principle. Thus, despite Davidson's efforts to effect a rapprochement with Freudian theory (Davidson 2004c), at its core

[22] There are a few complexities here, however, such as whether the module is better seen as *a*rational and not *ir*rational, and if the latter, whether it violates a 'basic' principle of irrationality (cf. p. 106ff.). I address these matters below.

[23] A noteworthy feature of the argument against the Competence Principle is that it appeals to a module embodying processes of wish-fulfilment which Freud regards as universal and non-pathological. One is perhaps inclined to regard pathological processes typically as matters not of competence but of performance-error, of interference with normal function. But this temptation does not gain a foothold in the case of the non-pathological processes that I have described. (Whether, however, pathological processes should always be viewed as performance-error is doubtful. Cf. below, p. 242 and n. 56.)

that theory must be seen as inimical—indeed, subversive—of a normativism like Davidson's.

More on Arguing Against the Threshold Principle

However, I wish to linger a bit over the argument against the Threshold Principle. The argument I have set out contrasts with another style of argument which can be made against that principle. W. E. Cooper (1980) advances an argument against Davidsonian charity based not on a scientific hypothesis but an imaginative thought-experiment (which he credits to Gilbert Harman).[24] He asks us to envision an individual, Napoleon, whose initially true belief-sets about cats and dogs gradually degrade, as from one day to the next, one of his beliefs about cats migrates to his belief-set about dogs and vice versa. Ultimately, he is left with nothing but false beliefs about cats and dogs. Both Cooper and Harman hold that normativists like Davidson and Quine, with their commitment to maximizing charity, must hold that the reference of 'cat' and 'dog' are exchanged just past the middle of the envisioned process. Thus, such charity can seem to rule out the possibility of a quite conceivable form of madness and so, Cooper concludes, should be rejected: "I do insist . . . that there *could be* such people whom we could recognize as being mad. And it would seem to count against a theory if it ruled out *a priori* this possibility" (1980, 39).[25] More generally, perhaps one can envision the truth of an individual's entire belief-set and the rational coherence among their propositional attitudes and

[24] In fact, Cooper takes the argument to undercut not just charity but materialism generally which he unwarrantedly takes to require charity.

[25] Cooper's argument levels specifically at maximizing charity, but he plainly holds that normativism rules out madness altogether ("there could be no madness" [1980, 39]). That conclusion, however, is much stronger than what his thought-experiment supports. At least some of the gap between his conclusion and this stronger claim can perhaps be filled by modifying his argument and re-directing it at the Threshold Principle instead of maximizing charity, as I do in the text.

behavior as gradually degrading over time, so much so that any arbitrary degree of falsity/irrationality would eventually be surpassed. If such ‘madness’ is a coherent possibility, then it can be taken as an argument against Davidson’s Threshold Principle. This sort of argument against the Threshold Principle is in an important respect the converse of that which I am emphasizing: whereas the latter begins with a state of predominant irrationality that subsequently yields to a state in which rationality predominates, the former begins with a state of predominant rationality that descends into irrationality.   I shall refer to these two sorts of arguments as *developmental* and *degenerative*, respectively. Although it is the developmental argument that I am urging against normativism, I shall consider the degenerative style of argument as well at a few points in the sequel.

With respect to the developmental argument, perhaps one may entertain doubts about the possibility of the id-module existing before other, more rational systems have come on-line. Indeed, Cavell expresses doubt that wishes can occur “prior to the formation of *some* beliefs and a considerable knowledge of reality.” “Concepts, and a knowledge of reality,” she maintains, are “as necessary a constituent” of desire as belief (1993, 166-67). In support, she appeals to a supposed conceptual connection between the concepts of desire and belief: “. . . the concepts presuppose each other. To desire that *p* is to be predisposed to act in a way that would bring *p* about, if one’s relevant beliefs about the world were true and there were no conflicting desires” (1993, 166-67). But whatever one thinks of this alleged conceptual connection, it is hard to see how it supports her conclusion that desire requires “considerable knowledge of reality.”[26] The phrase ‘relevant beliefs’, presumably, means something like ‘any beliefs one has about available means to the attainment of ‘*p*’. But unless the latter phrase is interpreted with existential import, Cavell’s alleged conceptual truth does not entail that desire requires any beliefs at all, let alone “considerable knowledge of reality.”[27]

Of course, it must be conceded to Cavell that concepts are a necessary constituent of desires, and one may wonder how the id, prior to the mind’s serious engagement with external reality, could come by the stock of concepts that figure in the unconscious wishes implicated in primary process. But the ready response is to deny the empiricist trend of the query: psychological theorists of recent decades have become comfortable with nativism with respect to both concepts (e.g., Fodor’s LOT) and beliefs (e.g., Chomskyean grammar), especially in modular contexts like the present one. Indeed, as the following passage illustrates, such nativism is eminently Freudian: “The content of the *Ucs*. may be compared with an aboriginal population in the mind. If inherited mental formations exist in the human being . . . these constitute the nucleus of the *Ucs*.” (Freud 1915, 195). There seem to be nothing amiss, then, in viewing the id-module as coming stocked with its own proprietary set of concepts

[26] Whether Cavell’s claim of the functional interdependence of belief and desire tells against the possibility of beliefs playing a part in the action-remote id-module is a separate matter, which I consider subsequently.

[27] Additionally, I note that even if large number of beliefs were required there would be a further step involved in concluding that they must be veridical. Note, further, that my reconstruction of primary process represented by the id-module *does* involve at least some beliefs, namely, those arising in the course of hallucinatory wish-fulfilling processes itself. However, I do not follow Cavell in taking beliefs to be *required* by wishes and desires.

pertaining to instinctual needs, etc.[28] So, I conclude, the bruited objections to the developmental argument fail.

Other Objections to My Argument

Other doubts, however, might be entertained with respect to my argument against the Competence and Compartment Principles. In the present section, I shall set out and attempt to rebut several possible objections.

Unconscious Sensation

First, I want briefly to address what might occur to one as an objection to the very possibility of wish-fulfilment. The potential problem arises from the fact that Freud conceives of infantile hallucinatory wish-fulfilment as unconscious.[29] Whereas dreams are conscious (or preconscious—that is, "capable of becoming conscious" [Freud 1915, 173]), the quasi-sensory experiences involved in infantile hallucinatory wish-fulfilment (and the pleasurable feeling of satisfaction to which they give rise) occur even before the infant develops a capacity for consciousness, in Freud's view. One might bridle at the possibility of unconscious imaginings, let alone unconscious pleasure.

But the existence of unconscious perception at least has become part of the stock-in-trade of mainstream cognitive psychology;[30] and if unconscious perceptions are a coherent possibility, then it is difficult to see how unconscious imaginings, states rather like perceptions only produced indogenously, are not. One might draw

[28] That is not to say that the id operates *only* with such concepts. Since its encapsulation is by no means complete on Freud's conception, as the psyche develops, its fund of concepts will be correspondingly enriched.

[29] Gardner (1993, 267n32) recognizes that "unconscious imagining" may "seem problematic," but he does not suggest a solution other than to hint that developmentally early imaginings may differ from later, more sophisticated ones.

[30] See, e.g., Kihlstrom (1999, 424-42) and Wilson (2002).

the line, however, at the existence of unconscious pleasure: Is pleasure not inherently qualitative and, therefore, essentially conscious? But if so, one acceptable response might be simply to suggest a slightly truncated structure for wish-fulfilment (at least of the infantile sort). In terms of Gardner's picture (see p. 166 above), this would mean omitting the feeling of satisfaction, (4), so that the wish-fulfilling representation, (3), leads directly to the termination of the wish, (5). Although this diverges from Freud's conception of such wish-fulfilment somewhat, I am not necessarily committed to defending every aspect of Freud's characterization of primary process.

But I am more inclined to defend a place for unconscious pleasure in wish-fulfilment. Indeed, in recent years many philosophers of mind have been attracted to a view of pain (and pleasure, presumably) that readily accommodates the possibility of unconscious pleasure. Without getting too deeply mired in the bog of theories of consciousness, one can say that on the present view a distinction is made between qualitative states like sensations, pain, and pleasure, and the *qualia* they possess when conscious. On this view, the essence of such qualitative states is their distinctive intentional content rather than any particular associated conscious *feel*.[31] Just what the nature of that distinctive content might be, and what is additionally required for such states to become conscious (and perhaps possess a quale), are questions which receive a variety of different answers from adherents of the view.[32] But when that additional element is lacking from particular episodes of such states, they will be

[31] In fact, the view is sometimes wedded to a sceptism about qualia. See, e.g., Rey (1997, 301-04).
[32] Peter Carruthers, e.g., following Michael Tye, takes pain to be a "perceived secondary quality of the body," which is phenomenally conscious only insofar as it becomes the non-inferential object of another, higher-order mental state. See, e.g., Carruthers (2004).

unconscious. Moreover, the same sorts of consideration that argue acceptance of unconscious perception can be marshaled in support of unconscious pain and pleasure. If the evidence supports the existence of states which, though unavailable to consciousness, substantially resemble conscious pain and pleasure with respect to their role in the psychic economy, then a good case can be made for viewing them as unconscious twins of their conscious counterparts.[33] So much by way of rendering the part played by unconscious pleasure in wish-fulfilment innocuous.

Is Basic Wish-Fulfilment Irrational?

A different potential problem for my argument is that one might wonder if the competence represented by basic wish-fulfilment is genuinely irrational. For, in the first place, inasmuch as wish-fulfilling processes appear rather efficient means for meeting needs of reducing psychic tensions, they may—*prima facie*—present an air of practical rationality. However, this appearance is deceptive. Perhaps it can be granted that mechanisms of wish-fulfilment manifest efficient, adaptive design. But to concede that is not to concede the phenomenon any relevant sort of rationality.[34] Indeed, because basic forms of wish-fulfilment as here rendered are automatic processes, in no wise to be construed as actions either overt or mental on the part of their bearer, they fall outside the sphere of practical rationality altogether.[35]

[33] On this point, cf. Carruthers (2004).

[34] At best it would support the rationality of some hypothetical designer who has the aim of constructing efficient, adaptive creatures. But the only relevant rationality is that of the bearer of the wish-fulfilling processes him-/herself.

[35] That is not to deny that the generation of beliefs (and other mental states) might sometimes possess the character of an intentional action and so fall within the scope of practical rationality (cf. the model of self-deception in Rey [1988], e.g.). Rather, it is simply to say the generation of beliefs (and hallucinatory experiences) in basic wish-fulfilment, on the Freudian conception reflected in the id-module, lacks an act-like character.

Of course, it might be *possible*—although textual support would be lacking—to offer a reconstruction of basic Freudian wish-fulfilment on which it turns out to be a species of intentional action: On such a reading, the id could be seen simply as electing to vividly imagine desired states-of-affairs when the real satisfaction of the relevant desires is barred. Basic wish-fulfilment would then be evaluable with respect to its practical rationality and might even plausibly turn out to be normatively rational. Indeed, it would be hard to find anything to reproach in such imaginative episodes when the infant (and perhaps even the young child) possesses virtually no capacity to engage in actions that might result in the genuine satisfaction of its real-world desires![36] But the suggested account of basic wish-fulfilment, although possibly of some interest in its own right, is simply not that which I am propounding for the purpose of making my anti-normativism argument.[37] Instead, I have offered as an empirical hypothesis an account on which wish-fulfilment dispenses with the character of intentional action and, thus, is not subject to evaluation by standards of practical rationality. On the present account, the problem wish-fulfilling processes pose for normativism rests, rather, on the breach of theoretical rationality they appear to commit.

But can the question perhaps be raised whether they do, in fact, violate canons of theoretical rationality? It is instructive, I think, at this point to compare basic wish-fulfilment with the sort of wishful thinking (WT, see p. 167 above) in which wishes

[36] I am not implying that phantasizing must be irrational for the older child or adult. However, for them, unlike the infant, serious question can at least arise whether, on given occasions, their time might be better spent in other activities.

[37] Nor, as noted, does it appear particularly Freudian. It will be seen, however, that Freud plainly assigns at least some role to intentional action in the generation of characteristic Freudian phenomena other than basic wish-fulfilment. I address this point at length in the next chapter.

immediately give rise to beliefs of identical propositional content without the interposition of any sort of hallucinatory experience. The theoretical irrationality of WT, in which beliefs are formed wholly at the prompting of wishes, is, I think, immediately apparent. However, when a hallucinatory experience is interpolated between wish and belief, the situation may appear relevantly different. For whereas the beliefs in WT clearly lack warrant, the beliefs in WF (wish-fulfilment) might seem to derive justification from the quasi-sensory experiences which precede their formation.[38] Certainly, in ordinary waking, conscious life, we regularly form what look like warranted beliefs on the basis of our sensory experiences. Might the beliefs in WF, though false, at least derive warrant from the quasi-sensory experiences that give rise to them and, therefore, count as rational?[39]

The problem with this line of thought is, I think, twofold. First, it oversimplifies what conscious beliefs formed on the basis of sensory experiences require for warrant. The presence of a sensory (or quasi-sensory) experience with a certain content does not always suffice to confer warrant on the corresponding perceptual judgments. Additionally, there may, for example, be constraints concerning how those judgments *cohere* with one's background beliefs.[40] To see this, suppose someone knowingly ingests LSD and hallucinates that there is a gorilla on

[38] At the very least, in the case of WF as opposed to WT there is what looks like a *reason* for one's belief, and perhaps a good one?

[39] Here and in the sequel I cease to distinguish between epistemic and rational probity. In fact, I am inclined to think that epistemology is properly subsumed under theoretical rationality. Against this view would perhaps count the fact that coherence among propositional attitudes (and behavior) is so often cited as the essence of rationality. For taking that as the hallmark of rationality would exclude epistemology from the purview of rationality, at least insofar as it concerns the formation of individual attitudes (beliefs) solely on the basis of non-attitudes (esp. sensation) or altogether without evidence but reliably (as seems to apply, e.g., on the Chomskyean account of innate tacit linguistic knowledge). In the present context, the issue is in any case moot, since my discussion of Chapter Two leaves no doubt that Davidson intends his charity-principles to concern the adherence to epistemic and rational norms indifferently.

[40] I am indebted to Georges Rey (personal communication) for this point and the illustration.

their shoulder. If they form the belief that there is a gorilla on their shoulder in those circumstances, then plainly their belief would lack warrant, precisely because it would fail to cohere with relevant background beliefs (about the effects of LSD, the likelihood of encountering a gorilla on the loose, etc.). But on the account sketched, wish-fulfilling beliefs are formed quite automatically on the basis of hallucinatory experiences, entirely without regard to coherence with background beliefs. So where these beliefs are formed against a background of beliefs with which they do not adequately cohere, they may similarly lack epistemic warrant.

"Fine," someone might object, "but the id-module does not clearly possess a fund of background beliefs that might raise problems of coherence in the formation of beliefs by the module." Perhaps, but this objection focuses too narrowly on the id-module itself. One form of basic wish-fulfilment, that constituted by undistorted dreaming, takes place in the presence of reality-oriented secondary process. So there will be abundant relevant background beliefs in the cognitive system considered as a whole. That those beliefs reside largely or wholly outside the id-module itself does not necessarily mean that they are irrelevant to processes of belief-formation within the id itself. As has been frequently noted, standards of ideal rationality[41] actually seem to require that in processes of practical and theoretical reasoning any and all relevant information held within the cognitive system as a whole be brought to bear. As Lisa Bortolotti observes, one reason a belief "might fail to be rational" is that it "is compartmentalized, that is, it does not cohere with other beliefs that belong to the

[41] And, of course, in the context of an attack on Davidisonian normativism, it is the ideal standards that matter.

same system and with the rest of the subject's behaviour" (2005, 199).[42] To the extent that dreaming wish-fulfilment flouts this normative requirement, its irrationality (on Davidson's very high standard) seems assured. It represents an irrational competence that should call into question Davidson's Competence Principle.[43]

However, some reservations may yet exist with respect to the 'developmental' argument that I am pressing against the Threshold Principle. For that argument requires envisioning the existence of the id-module prior to the development of reality-oriented belief-formation. But prior to that development, there will be no beliefs lodged in the broader system that the id-module will be required to consult in the course of generating its beliefs. In that case, the considerations which seemed to count against the rationality of dreaming wish-fulfilment appear to lapse with respect to the unconscious wish-fulfilment that is supposed to precede it developmentally. But then the developmental argument cannot get off the ground.

There is a fairly straightforward response to these reservations, however. Granted, such unconscious wish-fulfilment will not count as irrational in virtue of those considerations that tell specifically with respect to the developmentally later

[42] Granted, this point would lapse with respect to a polycentric normativism. But Davidson's normativism—the version at which my argument largely levels—is uncompromisingly monocentric (see p. 105, n. 139 above for the monocentric-polycentric distinction).

[43] A slight complication is that, as I discussed in Ch. 3, Davidson formulates his Competence Principle specifically with respect to 'basic' principles of rationality. Are the principles violated by the id-module basic ones in the relevant sense? This question perhaps admits of no answer, since, as discussed, it is difficult to formulate an acceptable sense of 'basic' in the context of Davidson's philosophy. In any case, the principles violated appear to be major rational principles. If a Davidsonian should care to deny their status as 'basic', I simply challenge him/her to do so in a way that avoids being *ad hoc*.

dreaming wish-fulfilment.[44] But unconscious wish-fulfilment would still seem to be irrational, namely, because the beliefs that are formed in the course of it are massively unreliable. Judged by the standards of externalist-reliabilist epistemology—which has acquired something of the status of orthodoxy in recent decades[45]—the belief-forming mechanism embodied in WF appears to fail miserably![46]

Of course, it might be objected that the mechanism consisting in forming beliefs based on what seems to be the case (unless there is reason not to), although resulting in abundant false beliefs in the case of wish-fulfilment, functions quite reliably for human beings when assessed with respect to their overall epistemic performance and, therefore, can bestow epistemic warrant even in the case of wish-fulfilment. This objection, however, assumes that the mechanism embodied in wish-fulfilment is the same one exhibited in conscious belief-formation. But, on its face, the former appears much more rigid in its operation; it seems to lack the qualified character (indicated by the phrase‘unless there is reason not to’ above) that its conscious counterpart possesses. Indeed, there is no compelling reason to suppose that the id-module, even if stocked with background beliefs, would have to be sensitive to them in the way that conscious perceptual belief is sensitive to salient

[44] It might be possible, however, to envision the cognitive system as congenitally stocked with various theories about the world (a folk physics, folk biology, etc.). Given the tendency of wishes to neglect real-world constraints, there would be a non-negligible possibility that beliefs formed by the id would conflict with such innate theories. In that case, the id could be convicted of irrationality for failing to check for consistency even prior to the development of secondary process. It must be conceded, however, that taking this argumentative tack would come at the cost of somewhat weakening the force of the developmental argument: To the extent that such innate theories are themselves consistent, both individually and collectively, they would seem to reduce the proportion of irrationality in the cognitive system overall. The id’s irrational operation would no longer be the *sole* factor to consider in judging whether the level of rationality of the system falls below that enshrined in the Threshold Principle. So I shall not press this tack.

[45] See Goldman (1979) for a defense of externalist-reliabilist epistemology.

[46] Indeed, aside from mechanisms that mimic reliable ones but—in a final fit of absurdity—reverse the sense of the beliefs formed, it is difficult to conceive a less reliable mechanism.

background beliefs.[47] So the assumption that wish-fulfilment is just a special case of the conscious mechanism is unjustified.

Moreover, even if it were one and the same mechanism at issue in both wish-fulfilment and conscious perceptual judgment, a case can be made that beliefs formed through processes of wish-fulfilment are still less than epistemically virtuous. For suppose that by chance an odd belief or two formed by an infant's wish-fulfilling module turned out to be true. One would hardly be entitled to attribute knowledge to the infant on the grounds that the beliefs in question are true and formed by a reliable mechanism. Thus, it appears to follow that the beliefs in question, indeed all infantile wish-fulfilling beliefs, lack justification. Since even infantile wish-fulfilment falls short of epistemological-rational probity, the developmental argument against the Threshold Principle can stand.

Is Basic Wish-Fulfilment Arational?

However, an opponent of my arguments may have one more trick up his sleeve. He might try insisting, not on the *ir*rationality of the processes on which the arguments are based, but on their *a*rationality, their exemption from evaluation by rational standards altogether.[48] But it is not clear on what basis a case for the arationality of those processes can be successfully made. The matter is complicated

[47] In fact, as I noted above (p. 170, n. 21), the id for Freud is not fully encapsulated. After secondary process begins to arise, Freud appears to see some contents as making their way from consciousness to unconscious. Though basic wish-fulfilment continues to operate (in the dreaming form, specifically), there is no indication that Freud sees wish-fulfilment as suddenly qualified in its operation by any sensitivity to background belief.

[48] Of course, in Ch. 3 I addressed an argument that sought to exclude modular processes from the scope of normativism because they are supposedly non-inferential (and, thus, not rationality-evaluable). Here I address the issue of their rationality-evaluability directly.

by the lack of an agreed-upon criterion of rationality-evaluability.[49] Of course, there are clear cases of rationality-evaluability (conscious decision-making and inference) and its lack (brute physical processes), but where between those extremes should one draw the dividing line? Of course, the line should be placed somewhere within the sphere of the mental, but perhaps the most straightforward suggestion—that rationality-evaluability is coterminous with the propositional attitudes—appears not wholly adequate, if only because the attitude of *entertaining* a thought seems exempt from assessment as to its rationality. But leaving this general question aside, what can be said for viewing processes like those of the id-module as arational?

There is, of course, a resemblance between wish-fulfilment and ordinary episodes of phantasy, and since the latter escape rational assessment, perhaps one might be tempted to think the former does as well. But that judgment ignores that wish-fulfilment on the present account differs from phantasy in involving belief in the states of affairs which are imagistically depicted.[50] The presence of beliefs would seem to bring wish-fulfilment into the ambit of rational assessment.[51] Perhaps, however, there is some pull towards saying that such beliefs, largely insulated from the wider stream of inference, decision-making, and action, should be viewed as harmlessly arational, not as noxiously irrational. But it is not clear that such innocuousness would remove the beliefs from rational assessibility. Rather, it would

[49] The question does not even seem to have garnered much attention. The very fact that I have had to coin the term ‘rationality-evaluable’ (on the analogy of ‘truth-evaluable’, which has some currency) to pick out one of the two important senses of the ambiguous ‘rational’, points up this lack of attention.

[50] Peter Carruthers (2006) presents an account which does implicate beliefs (or, at least, “belief-like” attitudes) in conscious phantasy, but the relevant attitudes are unconscious. On such an account, a question of the rationality-evaluability of phantasy arises, just as it does in the case of wish-fulfilment.

[51] Of course, one can resist this conclusion by denying that wish-fulfilment can (or should) be seen as involving beliefs or, indeed, propositional attitudes altogether. I consider this argumentative strategy below.

merely undercut the practical point of such assessment.[52] If wish-fulfilment involves belief, then that allows for their assessment by rational canons of reliability, coherence, etc. What could plausibly exempt them from that assessment? Moreover, the beliefs involved even seem to permit of a kind of reason-explanation couched in terms of the hallucinatory experiences that precede them: *a* comes to believe *p* because of his/her experience, where ‘because’ is understood as introducing a justificatory reason, not a purely explanatory one. And where there is reason-explanation, of course, there is rationality-evaluability.

Nor is it clear that the proponent of normativism *ultimately* gains much by denying the rationality-evaluability of wish-fulfilment, at least if the phenomenon is taken to involve any sort of propositional attitude at all. For the reasons which I have enumerated in Ch. 3, the restriction of the scope of normativism to a proper subset of propositional attitudes is highly problematic. The only half-way plausible maneuver available to a normativist (cf. p. 68 above) would be to hold that attribution of attitudes occurs in two phases, an initial round in which rationality-evaluable attitudes (and, especially, decision-theoretic ones) are attributed, and a second round, parasitic on the first, in which other attitudes are ascribed.[53] But the adherent of the parasitic strategy, then, confronts a dilemma. Either she can maintain that those attitudes are rationality-evaluable, in which case their irrationality tells against the Threshold Principle, or she can view them as parasitic on rationality-evaluable attitudes. But in the latter case, the parasitic strategy breaks down. For prior to the development of the

[52] Thus, though dreams, it appears, are not commonly assessed with respect to their rationality, this could simply be due to the lack of practical point in doing so. The attitudes involved in dreams are largely insulated from inference and action.

[53] Let us call this maneuver the ‘parasitic strategy’.

logical, reality-oriented processes of the ego, there will exist only those attitudes bound up with the id-module itself and, thus, no rationality-evaluable attitudes on whose basis the arational attitudes of the id can be attributed! So the scenario contained in the developmental argument, which on one characterization tells against Davidson's Threshold Principle, on another characterization tells equally against the parasitic strategy. Again, to the extent that wish-fulfilment involves propositional attitudes at all, there appears no way for the normativist to resist the force of my arguments against that position.

But, as mentioned, Gardner's account of basic wish-fulfilment (on which mine is largely based) differs from mine precisely in denying a place for propositional attitudes within that phenomenon. Accordingly, I shall consider what Gardner says on behalf of his characterization of wish-fulfilment and whether an alternative one which finds a place for propositional content can be defended.

A Place for Propositional Attitudes?

One quick argumentative route to the conclusion that wish-fulfilment—if a coherent phenomenon at all—must involve propositional content is simply to deny the possibility of non-propositional content. Adopting such a view might mean, for example, modifying Gardner's account of wish-fulfilment so as to interpret not only psychoanalytic wishes, but even the quasi-sensory experiences they cause as propositional (see, e.g., Rey 1997, 237-63). But though the view that all content is propositional has its defenders, I prefer not to stake my defense of a role for propositional content in wish-fulfilment substantially on this controversial premise.[54]

[54] Moreover, it's not clear that this approach by itself suffices to yield a conflict with Davidsonian Charity. For the propositional attitudes which on such an account constitute the quasi-sensory wish-

As noted above (p. 166), Gardner holds that it is unnecessary to introduce belief into basic forms of wish-fulfilment because, he thinks, the non-propositional sensory experience involved in them, (3), suffices to explain the latter phases of wish-fulfilment, namely, the feeling of satisfaction, (4), and termination of wish, (5). Moreover, Gardner implies that belief could not enter into infantile wish-fulfilment and dreaming because infantile mentality and sleep are conditions "where there is insulation from belief" (1993, 125). The suggestion is that, on the one hand, belief as a mental state has yet to develop in the young infant and, on the other, does not—at least at the outset—enter into mental activity during sleep.

Now Richard Wollheim offers a reconstruction of wish-fulfilment that differs from Gardner's in one important respect, namely, in seeing belief as an integral part of the phenomenon (Gardner 1993, 129-31; Wollheim 1979). As Gardner points out, for imaginings such as occur in wish-fulfilment to be "effective in gratifying or providing 'pseudo-satisfaction' for a desire [i.e., wish]," Wollheim holds that one must believe that what one imagines is the case (1993, 129). So Wollheim's view about what is required to explain the feeling of satisfaction and termination of wish in wish-fulfilment differs from Gardner's.[55] But we are all familiar with how a vivid fantasy (say, of an ice-cream cone on a sweltering day) affords pleasure and is able, at least temporarily, to alleviate an imperious desire (e.g., for relief from heat or thirst). It is able to do this without any obvious contribution from belief, inasmuch as one

fulfilling representation would seem to have a *subjective* content. That such attitudes should be caused directly by wishes does not seem to violate rational norms.

[55] Marcia Cavell's account of wish-fulfilment in (Cavell 1986, 495-507) derives largely from Wollheim's and agrees with his in seeing an attitude like belief as playing a role. She terms this attitude "proto-belief." I shall have more to say about Cavell's view below.

does not mistake the product of one's imagination for reality.[56] So why not similarly for wish-fulfilment? Thus, despite Wollheim, Gardner seems vindicated in holding that there is no *a priori* reason to introduce belief into wish-fulfilment.

But at least with respect to dreaming, there may be *a posteriori* reason of a rudimentary sort for accommodating belief in wish-fulfilment. First, there is the phenomenology of dreams: when one remembers one' dream of standing at a precipice fearing that one will fall over it, one seems to remember having *believed* that one was standing at a precipice. Moreover, dreaming frequently involves the experience of a variety of—sometimes quite strong—emotions (consider, for example, the emotions which linger after waking up after a sad dream or nightmare). Further, these emotions are at least often sensitive to the content of dreams in a way that seems to exhibit the same sort of rationality that waking emotions typically do. For example, the fear one experiences while dreaming of standing at a precipice seems commensurate to the situation of so standing. To make sense of this sort of emotional rationality in dreaming seems to require assuming that emotions experienced while dreaming are responses to propositional beliefs in dreaming, just as waking emotions are responses to waking beliefs.[57] So the familiar sort of dream would seem to involve beliefs. Thus, Gardner's claim that in sleep "there is insulation from belief" appears untenable. The upshot is that perhaps infantile hallucinatory wish-fulfilment need not involve belief, but to the extent that the undistorted dreams of young children at all resemble those of adults, then, the form of

[56] But do recall (see. p. 184 , n. 50 above) that at least one view about the processes involved in such conscious fantasies makes a place for unconscious "belief-like" states in explaining how such fantasies afford pleasure.

[57] Indeed, if a cognitive theory of emotions is true, emotions are *themselves* beliefs.

wish-fulfilment that they constitute would involve belief. Moreover, there seems no obvious conceptual bar to the possibility of either form of wish-fulfilment *potentially* including beliefs. Indeed, there seems no *a priori* reason even to rule out consideration of these forms of wish-fulfilment (propositional content and all) as serious empirical hypotheses.

As for the wishes involved in wish-fulfilment, as mentioned earlier, Gardner denies that they are ordinary wishes or desires, and, indeed, that they are propositional attitudes at all.[58] In essence, he seems to argue that they are not ordinary desires by pointing to their different functional role. Whereas ordinary desires are disposed to give rise to actions which aim to realize those desires' conditions of satisfaction, psychoanalytic wishes are disposed to cause sensory experiences which represent those wishes' objects as existing:[59] "Psychoanalytic wishes," he writes, "are necessarily engaged in the process of wish-fulfilment . . . which is not true of ordinary, conscious wishes" (1993, 126).

[58] Cavell (1986) agrees with Gardner that psychoanalytic wishes are not ordinary wishes, although she does apparently regard them as propositional attitudes.

[59] Some care, however, needs to be exercised with respect to the terms 'wish' and 'desire'. How clear is it, say, that ordinary *wishes*, as opposed to desires, are disposed to give rise to actions? At first sight at least, wishes do not appear to be decision-theoretically engaged in the way that desires typically seem to be. In his argument, Gardner does not draw a clear distinction between these sorts of attitudes. But there are certainly differences in the ways 'wish' and 'desire' are used in everyday speech, which suggests that what is true of the one sort of attitude cannot automatically be assumed to be true of the other. Thus, as Marcia Cavell observes, it is a "grammatical fact" that wishing, unlike desire, "can be counterfactual" (1993, 166); that is, one can wish—but not desire—for things one knows not to be the case. Cavell, however, seems to take this as evidence, not that wishes are wholly distinct from desires, but that they are a species of desire. She writes, "Wishes are typically desires upon which we might like to act but know we can or will not" (1993, 248n4). (Interestingly, Davidson himself repudiates such efforts to reduce kinds of propositional attitudes in this way—cf. p. 67 above.) Whatever can be said for this view, perhaps it explains Gardner's failure to clearly distinguish between wish and desire: If wishes are just desires we know we cannot act on, then they too could plausibly be seen as entailing a disposition to action, but one that perceived circumstances prevent from being engaged. (I assess Gardner's assertion that there is a functional difference between ordinary desires/wishes and psychoanalytic wishes below.)

Pre-Propositional Content

As for psychoanalytic wishes being pre-propositional, regrettably, Gardner provides no real defense of this claim. He merely invokes a distinction, which he has drawn earlier in connection with emotions. In that earlier discussion, he distinguishes between emotions which have particulars and those which have states of affairs as their formal objects: The objects of the former "are given by noun-phrases ('X hates a')," whereas those of the latter "by propositional expressions ('X hates its being the case that a is F')" (1993, 96). Citing the Ratman's supposed unconscious hatred of his father as an example, Gardner characterizes the former sort of emotion as "cruder" in form and, therefore, "unconditional," in that its non-propositional character precludes it from being responsive to justification or its absence (1993, 97). By making such a distinction credible with respect to emotions, he plainly hopes to render an analogous distinction between ordinary propositional wishes and psychoanalytic pre-propositional wishes plausible as well.

Whatever one thinks of pre-propositional content in general, the first thing to note by way of response to Gardner is that emotions and wishes are not the sorts of things to which pre-propositional content is commonly attributed. Usually, such content is appealed to in connection with perception and mental imagery, where what is intended is content that is altogether non-conceptual. In the case of emotions and wishes, however, Gardner seems to intend a form of content which, while non-propositional, *is* conceptual in character, although of a logically simpler sort than propositional content. To the extent that his proposal makes a quite novel use of conceptual content, we are justified in demanding of Gardner that he make a strong case for the theoretical indispensability of this use.[60]

Overall, though there appear to be some grounds for recognizing a category of emotions with non-propositional content, the case seems weaker with respect to wishes or desires. Let us first consider emotions. Unless one endorses a cognitive theory of emotions on which emotions are just certain kinds of beliefs, the appearance that emotions can have non-propositional content is not so easily explained away. With an emotion like fear, it may be possible to construe a sentence like 'I fear Bin Laden' in a given context as a mere pragmatic substitute for some longer sentence with a propositional object like 'I fear that Bin Laden could succeed in a terrorist plot'. Or perhaps one could propose a semantics on which a sentence like 'I fear Bin Laden' possesses a hidden logical form involving reference to some rather diffuse propositional content like *that Bin Laden will do harm to me or mine*.[61] However, sentences containing verbs like 'love' and 'hate', which refer to emotionally-tinged attitudes, seem to resist this kind of paraphrase. What suitable propositional content can one substitute for the object in a sentence like 'I love Greta Garbo'? Perhaps it will be possible to provide some sort of dispositional analysis of such sentences; perhaps, in fact, emotionally-tinged attitudes like love and hate may turn out to be dispositions to have a certain kind(s) of (non-dispositional) emotion towards an individual or thing. But until the project of providing such analyses meets with

[60] Granted, virtually every psychologist is committed to the existence of concepts. But this commitment falls far short of Gardner's notion that there are mental states whose intentional contents are bare concepts.

[61] I consider such semantic proposals in greater detail below.

success, the notion that some kinds of affective states have non-propositional content must be taken seriously.[62]

Even if there is a case to be made that some emotions have pre-propositional content, the case with respect to wishes appears weaker. Any sentence with a desire verb taking a noun-phrase as its direct object seems to permit paraphrase with a sentence whose object is propositional. For example, "The baby wants its bottle" is perhaps short for "The baby wants to have its bottle," which latter fails of having a propositional expression as object only because English grammar requires—approximately—that the subject of an object-clause be omitted when it would coincide with the subject of the sentence. Indeed, one fairly robust theory of the semantics of verbs like 'wants' (and even many emotion-verbs) interprets them as covertly embedding clausal complements as part of their logical form.

In general, there is a question of how to account for the semantics of what are called 'intensional transitive verbs' (ITVs), verbs like 'desire', 'fear', and 'look for' which, though—in at least some of their uses—taking noun phrases rather than clauses as their objects, create contexts which exhibit some or all of the typical marks of intensionality, namely, "substitution-resistance, the availability of unspecific readings, and existence-neutrality" (Forbes 2009). Thus, for example, the verb 'looks for' creates a context that resists substitution of co-referential expressions *salva veritate* and in which nouns and noun phrases carry no existential commitment. Again, in that context, quantified noun phrases permit of both specific and unspecific

[62] Note, however, that Gardner's claim that the lack of propositional content must render these affective states "unconditional" and unresponsive to justification is untenable. The best candidates for such states, love and hate, are—at least typically—as much grounded in reasons as any other sort of emotional state (cf. Soble [1989]).

readings: ‘I am looking for a wrench’ could mean either ‘There is some particular wrench which I am looking for’ or ‘I am looking for a wrench, but none in particular’. With respect to propositional-attitude verbs, there is an account of such phenomena available in terms of scope: a specific reading corresponds to a quantified noun clause having wide scope relative to the main verb and an unspecific reading to its having narrow scope. Similarly, an account of substitution-resistance and existence-neutrality might be framed in terms of the distinctive behavior of nouns and noun phrases when they occur in the scope of verbs. This has prompted some to try to explain the analogous phenomena with respect to ITVs in terms of scope as well. But since in standard first-order syntax quantified noun phrases require an open sentence as scope and, at first sight, suitable open sentences seem lacking for quantified noun phrases that appear as the objects of ITVs, some are drawn to a view (‘propositionalism’) on which ITVs covertly embed clausal complements. These complements provide the required scope for the quantified noun phrases.[63] The possibility of such a propositionalist account for desire verbs (and, even many emotion verbs[64]) significantly undercuts any support Gardner hopes to gain from linguistic idiom in making the case for a special category of pre-propositional desires or wishes.

Moreover, even if there is (or could conceivably be) some sort of wish with pre-propositional content, it is doubtful that Freudian wish-fulfilment ought to be

[63] Thus, for example, Forbes (2009, 7) presents the following as one propositionalist analysis of the sentence ‘Lois is looking for an extraterrestrial’: Lois is looking in order to make true the proposition that an extraterrestrial is such that she herself finds it.

[64] Though Forbes holds that a strong case for propositionalism can be made with respect to desire verbs, he cites various considerations that seem to count against propositionalism with respect to emotional verbs (Forbes 2009, 7-9). I do not think, however, that those considerations are decisive.

interpreted as limited to such wishes for at least two reasons. First, Freud accords an important place not just to libidinal instincts but also to ego-instincts in his psychology. Although the former might with—at least initial—plausibility appear to be directed at some particular object (or sort of object) which affords sexual pleasure broadly construed, the ego-instincts (for self-preservation and self-assertion in various forms) pretty clearly aim at states-of-affairs.[65] One seems forced to adopt the propositional idiom in formulating the wishes arising from such instincts. Further, in discussing children's undistorted dreams, Freud emphasizes that aside from the child's pressing biological needs (hunger, thirst, etc.), the major source of the child's dreams are desires that have remained unfulfilled during the previous day: "A child's dream," he writes, "is a reaction to an experience of the previous day, which has left behind it a regret, a longing, a wish that has not been dealt with. *The dream produces a direct, undisguised fulfillment of that wish*" (1916-17, 128). A wish "to go on the lake" which Freud cites shortly thereafter as "instigating" a dream evidently illustrates this source of dreams (1916-17, 129). But such ordinary wishes clearly have states-of-affairs as objects and, thus, further undercut Gardner's claim that the wishes implicated in wish-fulfilment are pre-propositional for Freud.[66] Accordingly, I feel that we can confidently assume that such wishes are propositional.[67]

[65] Thus, Freud (1916-17, 225) gives an example of a dreaming wish-fulfilment, albeit in an adult, which he maintains involves gratification of egoistic as well as libidinal wishes. He writes that a woman's dream of going to the theatre is a disguised wish-fulfilment in which "a satisfaction of her scopophilia was mixed with a satisfaction of her egoistic competitive sense", the latter particularly because her presence in the theatre symbolizes her beating her rivals to the marriage-altar.

[66] The claim is not necessarily that such ordinary desires receive wish-fulfilment *altogether* directly: left-over desires could give rise to wishes with the same propositional content which, in turn, produce the wish-fulfilling representations.

[67] Even if a good case could be made out that some of the relevant wishes are pre-propositional, Freudian theory itself seems to require that *some* be propositional; and, in any case, there seems no obvious reason why they *cannot* be.

But, ultimately, it is not imperative that I insist on a place for propositional wishes in wish-fulfilment. It suffices that beliefs be implicated in that phenomenon. For the irrationality of the id-module resides in the fact that beliefs are formed within it by an unreliable mechanism and without due regard to coherence with other beliefs within the cognitive system. The instigating wishes merely serve to *explain* those irrational beliefs; they do not strictly play a part in *constituting* the module's irrationality. It appears, then, that pre-propositional 'wishes' could take over the causal/explanatory role without detriment to my argument. Nor does Gardner adduce any considerations which militate against the phenomenon's including beliefs.[68]

Functional Role

But there remains Gardner's point that the wishes which receive wish-fulfilment cannot be ordinary desires inasmuch as they have a different functional role. Moreover, perhaps a similar point could be made with respect to the 'beliefs' that I have suggested must be supposed to be involved in dreaming wish-fulfilment inasmuch they would not have typical causal origins or exert influence on (overt) action. If both of these points are accepted, the result is a phenomenon comprising propositional attitudes but of an unfamiliar sort. Just this, in fact, is the picture of wish-fulfilment that Marcia Cavell presents in (1986) (cf. p. 187, n. 55 above). There she argues that the lack of action-engagement the propositional attitudes in wish-fulfilment exhibit precludes them from counting as ordinary desires, wishes, or beliefs. Instead, she characterizes these attitudes as 'proto-desire' and 'proto-belief'.

[68] Gardner's only comment in this regard (noted above, p. 166) is that beliefs are not *required* in an account of wish-fulfilment.

Would this impact the conflict between dreaming wish-fulfilment and Davidsonian Charity for which I have argued?

Well, first, it is not altogether clear that Gardner's sort of appeal to differing functional role establishes that the attitudes in wish-fulfilment cannot be the familiar ones. As mentioned, Gardner implicitly seems to assume that ordinary wishes are just desires accompanied by the belief that the object of one's desire is not (readily) attainable. This view possesses a certain plausibility, and if true, it would mean that Freudian wishes cannot just be ordinary wishes (appropriate desire-belief pairs) since the id cannot easily be supposed to possess such reality-based beliefs prior to the development of secondary process. But this view of ordinary wishes requires defense, of course. As noted, the decision-theoretic engagement of ordinary wishes cannot simply be assumed. If they in fact lack it, then much of the supposed functional difference between ordinary wishes and Freudian wishes disappears. Moreover, it is not a given that ordinary wishes do not equally give rise to imagistic representations of the wishes' satisfaction (if not the corresponding beliefs, as in my account of wish-fulfilment), although such imagery would certainly be more subject to voluntary control than that bound up in Freudian wish-fulfilment. In that case, the supposed functional difference between ordinary and Freudian wishes might seem to lapse altogether.

Moreover, Gardner's appeal seems to rest on a particular view of the semantics of mental vocabulary which Georges Rey calls '*a priori* functionalism' (Rey 1997, 186-87). This is a variety of functionalism which takes the specification of the relations constitutive of mental states to be conceptual truths and, thus, knowable *a priori*. It contrasts with what Rey terms ‘psychofunctionalism’, which views mental states as natural kinds whose causal roles are scientific essences to be discovered *a posteriori* (1997, 187-89). On this latter view, in advance of investigation, it seems possible that the best psychological theory could turn out to be one where one and the same kind of mental state (wishes, say) could have a dual life, as it were, sometimes exhibiting the familiar causal relations of wishes, sometimes those of the desiderative element in wish-fulfilment. More precisely, it might belong to the causal role of wishes, say, to behave in familiar ways in one set of circumstances (when conscious, e.g.) and in unfamiliar ways in a different set of circumstances (when unconscious, etc.). If psychofunctionalism is the correct view of the semantics of mental terms, then, psychoanalytic wishes and beliefs could turn out to be just the ordinary ones, despite any *prima facie* functional differences between them and their familiar counterparts.

Again, it is not a given that beliefs, wishes, or even desires, will turn out to have the sort of essential relations to outward action that both Gardner and Cavell take for granted in arguing for the distinctness of ordinary attitudes from Freudian ones. Georges Rey describes an approach to functionalism, ‘molecular functionalism’, which allows that particular kinds of mental states might be defined without reference to action (1997, 194-96). On a more standard ‘holistic’ functionalism, mental states are typically defined all at once with respect to their functional role in a network of causal relations encompassing all kinds of mental states, as well as behaviors, physiological states, etc. But molecular functionalism permits kinds of mental states to be defined in smaller clusters (‘molecules’), perhaps with those clusters themselves organized hierarchically within cognitive systems. As Rey points out, one advantage of this sort of functionalism is that it accommodates the intuition that a system could lack some familiar kind(s) of mental state and still count as possessing a mind.[69] More pertinently, as Rey observes, “it finally affords us the possibility of completely freeing the identification of a psychological state from any tight connection with behavior. It’s perfectly open to the molecularist to suppose . . . that few (if any) of the states of subsystems of the mind are actually defined in terms of actual behavior” (1997, 195). But in that case one cannot rely, as Gardner and Cavell do, on psychoanalytic attitudes’ apparent lack of a (direct) link to overt action to argue their distinctness from ordinary beliefs, wishes, and desires. Indeed, the defining functional features of these states could turn out to be entirely internal.

But even if they are better construed as distinct types of attitudes, this would not automatically cancel the irrationality of dreaming wish-fulfilment comprising such merely belief-*like* and wish-*like* propositional attitudes. Although matters are a bit murky at this point, it is not implausible to think that any attitudes with the respective directions of fit of wishes and beliefs inserted into the structure of mental states constitutive of wish-fulfilment would yield an irrational structure. So we do not necessarily need to resist Gardner’s and Cavell’s claims that the states involved are novel types to sustain the conflict between wish-fulfilment and Davidsonian Charity.

[69] Thus, Rey writes, “e.g. people who can’t feel pain, or psychopaths who don’t form the usual personal attachments, can surely have beliefs and desires and other perceptions” (1997, 194). As Rey points out, if states are hierarchically organized, those at the top of the hierarchy could drop out while leaving the rest intact.

But this makes Cavell's stance towards wish-fulfilment in (1986) somewhat puzzling. Cavell's philosophy of mind is that of a committed Davidsonian.[70] Moreover, in her article she offers her account of wish-fulfilment not just as an interpretation of Freud but as account of what she plainly takes to be a real phenomenon (1986, 496). But, as I have suggested, even on her own interpretation that phenomenon fundamentally conflicts with Davidsonian Charity! Moreover, her account evidences a willingness to countenance the attribution of certain kinds of propositional attitudes even to infants, which flies in the face of Davidson's exclusion of infants from that form of intentionality. Perhaps Cavell thinks she avoids conflict with Davidson because she takes him to limit the scope of his interpretavism (and Charity) to just familiar propositional attitudes. But as I suggested above (p. 68), he apparently takes himself to be offering a general account of propositional attitudes. Granted, Cavell can escape conflict by narrowing the scope of Davidsonian interpretavism. But, as I have argued in Ch. 3, diminishing the ambitions of that theory entails a corresponding diminishment of its interest—and plausibility.[71] In fact, one might expect a Davidsonian like Cavell to opt for a view of wish-fulfilment like Gardner's which denies that basic forms of it involve propositional content at all. But, of course, as I have argued, this tack itself—ultimately—fails to dissipate the conflict with Davidsonian Charity.

[70] Although Wittgenstein has also influenced her thinking (cf. Cavell 1993).

[71] As I pointed out, it seems unlikely that there should be two entirely disjoint constitutive bases for propositional content. And to the extent that *some* suitable non-interpretavist basis would need to be sought for the unfamiliar propositional attitudes, Davidsonians would face the—perhaps recalcitrant—problem of explaining why that basis is unsuited to serve for the familiar propositional attitudes as well. So a limitation of the scope of interpretavism such as Cavell seems to require renders it unsatisfyingly narrow, implausible, and unstable.

In sum, my argument against Davidsonian normativism, based on the scientific possibility of a module embodying processes of basic Freudian wish-fulfilment, withstands a number of objections that might possibly be raised against it. This argument should itself suffice to raise substantial doubt about the propriety of normativist claims that typical charity principles constitute *a priori* constraints on mental agency. However, in the following chapter I widen out my argument to take in *non*-basic forms of Freudian wish-fulfilment as well as some other characteristic Freudian phenomena.

Chapter Five: An Argument From General Freudian Phenomena

Introduction

Although my argument thus far has focused only on basic forms of wish-fulfilment, it is interesting to consider whether an analogue of my argument can be run with respect to a broader array of Freudian phenomena. Whether it can or not will depend on just how those phenomena are to be understood. Accordingly, I consider which of several possible accounts of those phenomena appears to be Freud's own, and whether when so interpreted, the relevant phenomena put pressure on normativist principles. My focus shall be largely, though not exclusively, on the various less basic forms of wish-fulfilment which according to Freud develop subsequently to the infantile forms previously described. Key interpretative questions which I address include the following: whether Freudian phenomena possess an act-like character, that is, whether at bottom they are actions, or the causal consequences of actions, performed by the unconscious; and in precisely what sense neurotic symptoms, parapraxes, etc., constitute wish-fulfilments for Freud.

Freudian Phenomena

Roughly, the psychological phenomena at the heart of Freudian theory divide into two classes: (1) defense mechanisms such as repression and resistance that function to make (or keep) contents unconscious, and (2) the various forms of wish-fulfilment, both basic and non-basic, that arise as outlets for defended-against contents.[1]

Defense Mechanisms

Of the defense mechanisms, repression is primary. Through it, wishes which are unacceptable to consciousness because they conflict with the moral standards incorporated into the superego are consigned by the ego to the unconscious. Moreover, memories of traumatic experiences or memories which are merely associated with traumatic experiences (and, therefore, run the risk of evoking traumatic memories) are subject to repression. Resistance, in turn, occurs when the ego instigates behaviors that serve to fend off forces that threaten to raise repressed contents to consciousness. In particular, a patient in treatment may act precisely so as to *sabotage* that very treatment.[2] How repression and, especially, resistance might operate in detail will occupy us subsequently.

Characterizing Forms of Wish-Fulfilment

Adequately characterizing the various forms of wish-fulfilment requires some length. Aside from infantile hallucinatory wish-fulfilment and the undistorted dreams of young children, Freud counts as forms of wish-fulfilment such varied phenomena as the distorted dreams of older children and adults, parapraxes, jokes, and neurotic symptoms like the bodily manifestations of conversion-hysteria and the obsessive

[1] Strictly, because the various dreamwork processes serve a censoring function, they should be reckoned among defense mechanisms. But because of their central role in non-basic wish-fulfilments, I treat them under the latter heading.
The classification given is rough partly because Freud describes some phenomena involving wishes that more resemble straightforward satisfactions than wish-fulfilments. Thus, for example, as I suggest below, Freud sees the unconscious as sometimes helping itself to realistic means in order to achieve its ends.

[2] So described, the irrationality of such resistance appears patent. But I postpone consideration of the irrationality of this and the other Freudian phenomena.

thoughts and symptomatic acts of obsessive-compulsion. What is peculiar to these forms of wish-fulfilment is that, in some sense, they wear a disguise.

Freud famously wrote that “dreams are the royal road to the unconscious,” by which he meant that an appreciation of the distinctive contents and modes of operation of the unconscious is most readily gleaned via consideration of the character of dreams. And it is with respect to dreams that Freud expounds in greatest detail the nature of disguised wish-fulfilments. As the child grows, the infantile wishes of the id become unacceptable to consciousness and are subject to the repressing forces of the ego and superego. Accordingly, the possibility of a straightforward satisfaction for them through action or even of an undisguised ‘gratification’ of them through dreaming wish-fulfilment is closed off to them. Hence, according to Freud, they find their way to a less direct expression: “. . . we can say of an infantile dream that it is the open fulfilment of a permitted wish, and of an ordinary distorted dream that it is the *disguised* fulfilment of a repressed wish . . .” (1916-17, 217). This disguise is effected through the processes collectively known as the ‘dreamwork’. The dreamwork consists of the various processes by which the wishes instigating dreams in adults (the dream’s ‘latent content’) are rendered unrecognizable in the dream-imagery which expresses them (‘manifest content’). These processes most centrally include condensation and displacement, whose operation (cf. p. 164 above) Freud sees as a characteristic feature of the id.

By ‘condensation’, Freud understands “the fact that the manifest dream has a smaller content than the latent one, and is thus an abbreviated translation of it” (1916-17, 171). Especially, it refers to the fact that “latent elements which have something in common" are "combined and fused into a single unity in the manifest dream" (*ibid.*). Thus, for example, "different people" in the latent content may in the manifest content be "condensed into a single one." As Freud notes, such fusion makes the original sources unrecognizable (thereby serving the function of dream-censorship), rather in the way that taking "several photographs on the same plate" makes the resulting image "blurred and vague" (1916-17, 172). By 'displacement', Freud understands the process in which "a latent element is replaced not by a component part of itself but by something more remote" and, secondly, "the psychical accent is shifted from an important element on to another which is unimportant, so that the dream appears differently centred and strange" (1916-17, 174). With respect to the former, the item replacing the latent element (or elements—cf. condensation) will be one that bears some sort of similarity or other associative connection, however tenuous, with the replaced item.[3] The creations of the dreamwork, then, rendered unrecognizable in their import by such processes, are able to escape the repressing forces of ego and superego and enter consciousness undisturbed.

The pattern which Freud describes with respect to dreams, of course, he claims to discern in a number of other psychic phenomena as well. Thus, a neurotic symptom (or at least particular features of it) is supposed to be a manifestation of latent unconscious thoughts, though the link is rendered unrecognizable through the distorting influence of dreamwork-like processes. Thus, for example, Freud describes an obsessive-compulsive patient, a woman of about thirty, who, Freud writes,

[3] In some instances, the replacing element will be drawn from a relatively stable stock of "symbols," the common inheritance of humanity (or a culture, at least); in others, they will constitute more private associations. In either case, however, it is essential that the "meaning" of the replacing element be opaque to consciousness. Indeed, the link between replacing and replaced element may be so remote as to depend on a mere punning, verbal connection.

> “performed (among others) the following remarkable obsessional action many times a day. She ran from her room into another neighbouring one, took up a particular position there beside a table that stood in the middle [with a prominent stain on its tablecloth], rang the bell for her housemaid, sent her on some indifferent errand or let her go without one, and then ran back into her own room.” (1916-17, 261)

Freud connects the symptomatic act to an incident on the woman’s wedding-night about ten year’s prior. Her husband had been impotent, but not wishing to be ashamed before the chambermaid when she made the bed, he had poured ink on the sheet, “but not on the exact place where a stain would have been appropriate” (262). On Freud’s interpretation, the woman’s symptomatic act essentially repeats the traumatic scene, only in such a way as to correct it: in the obsessional action, “the stain is in the right place,” and, thus, “she was also correcting the other thing, which had been so distressing that night . . . his [i.e., her husband’s] impotence” (262-63). “[I]n the manner of a dream,” the symptomatic act represents as fulfilled an unconscious wish on the part of the woman that her husband had not been impotent. As in a dream, the origins of the act in the woman’s traumatic memory and unconscious wish, to which it bears such significant analogies, is concealed by the distorting influence of the dreamwork. By displacement, elements of the memory and wish are replaced with items with which they are associatively or symbolically connected, for example, the bed with the table (the latter being, according to Freud, a regular symbol for the former) (262).[4] Most crucially, Freud maintains, “This symptom was fundamentally a wish-fulfilment, just like a dream . . .” (299). Similar remarks apply to less pathological phenomena such as jokes and parapraxes.

[4] The concept of displacement constitutes something of a genus for Freud. In fact, Freud identifies a number of more specific processes by which unconscious contents are rendered unrecognizable in dreams, symptoms, etc. On these various processes, see Suppes and Warren (1980).

A Clinical Portrait

To get a fuller appreciation of Freudian phenomena, however, it will be helpful to describe in a little detail the clinical picture presented to Freud by a case of a nineteen-year old girl suffering from a combination of agoraphobia and obsessional neurosis (1916-17, 264-69). In his description of the case, Freud mentions that she had become very irritable, especially towards her mother, depressed, given to indecision and doubts, and unable to cross wide streets by herself. But Freud dwells particularly on a "sleep-ceremonial" the girl had developed, to the consternation of her parents. On the pretext that she needed quiet, she required that all clocks in her bedroom be stopped or removed, and that all flower-posts and vases be collected so they would not disturb her sleep by falling over and breaking. She herself admitted how feeble a justification her need for quiet provided for these measures. Moreover, that she required the door to the hallway between her room and her parents' to remain open altogether contradicted her alleged motive. Other aspects of her sleep-ritual concerned her bed. She demanded that her pillows not touch the bedstead, and that the smaller topmost pillow form a diamond against that below. On this pillow, her head had to be placed along the center-line of the diamond. Moreover, she would take pains to collect the feathers of her eiderdown at one end before placing it on her bed; but then she would immediately smooth out the feathers that had so accumulated. While performing these rituals, she would be assailed by doubts that she had done them successfully and be forced to repeat them. As a result, an hour or two would pass in which she remained awake and kept her parents from sleeping as well. Freud's initial efforts to interpret the patient's symptoms were met with resistance: "I was obliged to give the girl hints and propose interpretations, which were always rejected with a decided ‘no’ or accepted with contemptuous doubt” (266). Gradually, however, the girl came to wholly accept Freud’s interpretations and, in proportion as she did, her symptoms disappeared.

Freud interprets the girl’s removal of clocks to ensure her sleep thus: The girl had been disturbed by repeatedly having been awakened with a throbbing in her clitoris from sexual excitement. When the symbolic relationship between ticking clock and throbbing clitoris is seen, then her insistence upon removing the clocks is recognizable as an expression of her fear of being so awakened. Her precautions with regard to flower-pots and vases is partly explained through the fact that “flower-pots and vases, like all vessels” are “female symbols” (267). Moreover, free-association to the ritual led the girl to recall an incident in which she had cut herself on a vase and bled profusely, which memory led her, in turn, to recall her anxiety at a later date that “on her wedding-night she would not bleed and would thus fail to show that she was a virgin” (267). Thus, Freud interprets her “precautions against vases being broken” as “a repudiation of the whole complex concerned with virginity and bleeding at the first intercourse—a repudiation equally of the fear of bleeding and of the contrary fear of not bleeding” (267).

As for the patient’s rule that pillow and headboard not touch: the girl admits that the pillow had the meaning of a woman for her, the headboard that of a man. So Freud interprets the rule as a means “by magic” to keep her mother and father apart, that is, from having sex. This interpretation is strengthened by the fact that she remembered as a child having achieved this aim more directly. First, she had simulated fear so that the door between nursery and parents’ room would remain open (an arrangement preserved in her sleep-ceremony). Later, she would be permitted to lie between her parents in their bed, and even to take the place of her mother when she grew too big to be comfortably accommodated between them.

The patient's gathering of feathers in the eiderdown (like pillows, a female symbol) represents her mother as pregnant, while the subsequent smoothing of it out Freud interprets in terms of the girl's long-standing fear of her mother becoming pregnant with a little competitor to herself, as it were. As for her placing her head along the center-line of the diamond formed by small pillow atop large pillow: The small pillow stands for the daughter, the large pillow for the mother. The diamond serves as a universal symbol of female genitalia; the head, of the penis. So this aspect of the ritual represents her "playing the man and replacing the male organ by her head" (268).

While noting that several distinct phantasies underlie the girl's symptoms, Freud finds their "nodal point" in the girl's erotic attachment to her father: In short, the girl was suffering from an Elektra complex. He speculates, further, that this may explain the girl's hostility towards her mother. I shall recur to the particulars of this case and Freud's reading of them in the sequel.[5]

Interpretative Issues

Various interpretative issues arise when one considers Freudian phenomena like those exhibited in the preceding case. In the present section, I consider these issues in detail.

[5] I shall refer to the girl as Elektra.

Intentionalism

Probably the most fundamental question is what role Freud assigns to *intention* in the production of these phenomena. As several commentators point out (see, e.g., [Gardner 1994] and [Smith 1999]), there is a reading of Freud that views such manifestations on the model of the practical syllogism. On this view, a symptom, say, is an intentional action; it is performed for an (unconscious) reason, namely, because one (unconsciously) desires *p* and believes that the symptom is a means to achieve *p*. Following Smith, I shall refer to this view as 'intentionalism'. [6] In fact, adherents of intentionalism might be tempted to see any or all stages of the process by which on Freud's account symptoms, etc., are produced and sustained in such terms: (1) the generation of symptoms and other outward manifestations, (2) the generation of wish-fulfilling imagery and dreams, (3) the repressions which consign wishes (and other unacceptable contents) to the unconscious and, thereby, preclude their straightforward satisfaction, and (4) the resistance which wards off forces that threaten to bring those unconscious contents to consciousness.

The intentionalist reading of Freud has sometimes found favor among those committed to a hermeneuticist approach to the mind (e.g., Flew [1956] and Toulmin [1954]). The advantage of this reading for the hermeneuticist, of course, is that it assimilates the pattern of psychoanalytic explanation to the rationalizing pattern which they see as at the heart of folk psychology, and, thus, it supports the

[6] Intentionalism as here understood requires *unconscious* reasons which rationalize symptoms, parapraxes, etc. Many will hold that symptomatic acts, like the obsessive-compulsive's rituals, permit reason-explanation in terms of an agent's conscious desires and beliefs. Elektra, for example, justifies some of her rituals in terms of her desire not to be awakened during the night. Perhaps she would justify others simply in terms of an intrinsic desire to perform them. It cannot merely be taken for granted that explanations of her acts in such terms are not genuine reason-explanations of them. (Whether they are or not raises complex issues that I explore below [see p. 217, n. 17].) But if they are genuine reason-explanations, they are not the sort of unconscious reason an intentionalist account of such phenomena requires.

methodological and ontological subsumption of psychoanalysis under folk psychology. Particularly relevant is the variety of intentionalism encouraged by Davidson's reading of Freud (Davidson 2004c) which sees symptoms (*qua* species of irrationality)[7] as the effect of interacting compartments. Though Davidson conceives of the interaction itself as brute causal (cf. p. 114 above), effects within one compartment relate logically to their distal causes (beliefs and desires) in another compartment as actions do to the reasons that rationalize them. So a Davidsonian reading of Freud can be seen as a kind of intentionalism.

Was Freud an intentionalist? It is difficult to judge with any confidence the precise extent of Freud's application of the rationalizing pattern of explanation. Cavell (1993, 180), for example, emphasizes that Freud's "last revisionary work on repression and related matters—*Inhibitions, Symptoms and Anxiety* (1926)—builds squarely on a reason-explanation model." But even earlier there are unmistakable intentionalist elements in Freud's theory. Thus, Freud describes the case of a middle-aged woman suffering from delusions of jealousy. One day she divulges to her housemaid that "The most dreadful thing that could happen to me would be if I were to learn that my dear husband was having an affair . . ." (1916-17, 249). The next day she received a note in which her husband was accused of having an affair with a girl whom the maid hated. Although the woman immediately saw through the ruse, from that day forward her confidence in her husband's fidelity was shaken, despite the fact that she had no good reason whatsoever to suspect her husband. Freud interprets the

[7] In fact, whether and how symptoms, etc., should be seen as irrational requires careful consideration. In principle, an intentionalist reading of Freud might seem to open up the possibility of a *hyper-rational* interpretation of Freud on which Freudian phenomena actually turn out, not just rationalizable, but normatively rational. More on this subsequently.

woman's delusion of jealousy so: Unconsciously, the woman is in love with her son-in-law, a fact which occasions considerable guilt in the woman. To assuage this guilt, she contrives by means of the housemaid to obtain evidence, however flimsy, to support a delusional belief in her husband's infidelity. On Freud's reading, then, the woman's initial revelation of her fear to the housemaid must be seen on the model of the practical syllogism: she unconsciously desires to obtain evidence of her husband's infidelity and undertakes a means to its attainment.[8]

Another clear illustration of Freudian intentionalism arises in connection with the tablecloth-lady (see p. 204ff. above). Freud interprets the symptoms she exhibits (including her ritual involving the tablecloth) thus: "By means of her symptoms she continued to carry on her dealings with her husband. We learnt to understand the voices that pleaded for him, that excused him, that put him on a pedestal and that lamented his loss" (1916-17, 273). In particular, Freud writes, "Although she was young and desirable to other men, she had taken every precaution, real and imaginary (magical), to remain faithful to him." The 'real' precautions referred to include the fact that "[s]he did not show herself to strangers and she neglected her personal appearance . . . ." Here again one see means straightforwardly employed for the attainment of a desired end on the model of the practical syllogism.[9]

[8] This is not to say that it would be impossible in principle to hypothesize some non-intentionalistic *mechanism* responsible for the woman's behavior. But, as in this case, where the intentionalist reading appears most natural and Freud gives no indication of repudiating it, I think it can safely be ascribed to him.

Note, by the way, that it is only the woman's revelation of her fear to the housemaid that I am addressing here. I do not consider, for example, whether Freud would explain the woman's delusion of jealousy itself intentionalistically. Nor do I consider such questions as why an unconscious sense of guilt should require relief, why producing the delusion demands that evidence be manufactured, or how such flimsy evidence forms a sufficient basis on which to rest the delusion.

[9] Perhaps one can also see the aspect of Elektra's sleep-ritual (p. 206ff. above) whereby she insists that the door to the hallway between her room and her parents' remain open as a realistic means to achieve

But there is another, somewhat more complicated feature of the case of the tablecloth-lady that seem to fit the intentionalistic pattern as well. In particular, Freud maintains that “the deepest secret of her illness was that by means of it she protected her husband from malicious gossip, justified her separation from him and enabled him to lead a comfortable separate life” (263). Freud’s point is that the woman, although “struggling with an intention to obtain a legal divorce” because of her husband’s impotence, uses the illness itself to provide her husband a reason to separate from her in a way that allows him to save face. Moreover, with respect to Elektra, Freud writes, “We . . . may suspect that she had become so ill in order not to have to marry and in order to remain with her father” (273). In each case, Freud posits a clear means-end relation between the patient’s illness and a supposedly desired end.

Such intentionalism necessarily complicates the picture of the causation of the symptoms of illness. I shall turn to the question shortly how Freud conceives of the causation of symptoms in detail. But this much is clear: by and large he construes them as wish-fulfilments. Thus, as we have seen, Freud offers interpretations of the rituals of the tablecloth-lady and of Elektra that view them as substitute-satisfactions for unconscious wishes or desires. So, for example, Elektra’s insistence that her pillow not touch the headboard should be seen as a wish-fulfilment of her desire to keep her parents apart. But inasmuch as Freud sees an intention to remain with her father behind her illness as a whole, that intention must also becomes part of the

her supposed aim of preventing her parents from having sex. A complication in this case, however, is the fact that Freud links this aspect of her ritual to her having feigned fear as a young child so that the door between nursery and parents’ room would remain open. The associative connection with the earlier behavior may suggest that the later ritual has more the character of a symbolic, wish-fulfilling act than of a measure undertaken to realistically achieve an end.

causal story of her symptom. So the intentionalism Freud posits with respect to the illness as a whole would seem to entail at least a measure of intentionalism with respect to symptoms themselves.[10]

Anti-Intentionalism

Yet many more-recent commentators interpret Freud's account of symptoms and other wish-fulfilments non-intentionalistically. Thus, for example, Gardner writes that "the true form of psychoanalytic explanation makes a clean break with practical reasoning" (1994, 497). Again, endorsing 'anti-intentionalism', Smith interprets parapraxes, symptoms, etc., as mere "manifestations of wishes," whereby he intends "a mode of causation less constrained by the content of the wish than the conclusion of a practical syllogism involving the wish as a premise" (1999, 173).[11] Aside from the views of commentators, there's the fact that the trend of Freud's discussion of the dream-work is to see it as possessing the character of *process* not action; and with respect to symptoms specifically, he writes, ". . . dealing with a conflict by forming symptoms is after all an automatic process . . ." (1916-17, 385).[12]

Moreover, problems arise when one tries to fit symptom-formation, etc., into the mold of the practical syllogism. It requires attributing to the unconscious truly bizarre beliefs to the effect either that a given symptom constitutes the attainment of

[10] Things quickly become murky when one asks how the causal contributions of the two desires relate, the desire behind the illness as a whole and the desire behind the symptom as wish-fulfilment. Do they over-determine the symptom? Do they both enter intentionalistically into a process of practical-reasoning that simultaneously seeks to "kill two birds with one stone," produce an illness and afford substitute-satisfaction? Or does the intention to produce illness merely avail itself of non-intentionalistic wish-fulfilling psychic mechanisms in order to achieve its end, as if it were turning on machines? This last model of the production of symptoms especially rather interestingly combines intentionalist and non-intentionalist elements. Yet another model might see the intention as taking over responsibility for the symptoms when the automatic mechanisms that initially generated them begin, for whatever reason, to give out.

[11] Compare also Cavell (2002) on defense mechanisms: "Repression, along with more specific defense mechanisms like displacement, consists of purposive, quasi-automatic, nonintentional mental acts."

[12] In this context, it is worth noting the existence of efforts to model such processes in non-intentionalist computational terms (cf. p. 170, n. 20 above).

an object of (libidinal) wish or serves as a means to its attainment. Thus, in connection with Freud’s view of paranoia as a disguised wish-fulfilment for homosexual desires, Adolph Grünbaum writes,

> In the psychoanalytic explanation of a paranoiac’s delusional conduct, can the afflicted agent be warrantedly held to have “reasons” for his/her behavior such that he/she unconsciously believes it to be a means of attaining the fulfillment of his/her homosexual longings? Can the paranoiac be warrantedly said to have unconsciously *intended* his delusional persecutory thoughts and comportment to accomplish his erotic objectives? (Grünbaum 1984, 76-77)

The point Grünbaum highlights is not that it is impossible that the unconscious might possess such bizarre means-end beliefs, but merely that one would need strong evidence to attribute such beliefs. However, as Gardner notes, an intentionalistic account confronts a major explanatory burden: “Where do such irrational desires and beliefs come from, and why are they not integrated into, and so dissolved away by rational mental functioning?” (1994, 497).

Nonetheless, even with respect to wish-fulfilling symptoms, Freud sometimes uses what might appear an intentionalistic idiom. Thus, of Elektra’s requirement that her pillow not touch the headboard, Freud writes, “she wanted—by magic, we must interpolate—to keep the man and woman apart—that is, to separate her parents from each other, not to allow them to have sexual intercourse” (1916-17, 267). Again, Freud writes of the tablecloth-lady that “she had taken every precaution, real and imaginary (magical), to remain faithful” to her husband (273). Taken at face value, Freud’s phraseology suggests, in the case of Elektra, that she separates pillow and headboard as a magical or imaginary means to achieve the end of separating her parents.[13]

Because the actions undertaken are not realistic means to the desired ends, Freud qualifies them as ‘magical’ or ‘imaginary’. But question arises as to what these adjectives signify in this context. Whatever else magic as it has been practiced in various cultures includes, presumably it involves *beliefs* that particular actions are effective for achieving specific ends that appear unfounded from a scientific or broadly naturalistic standpoint. So perhaps Freud can actually be taken as attributing to Elektra the absurd belief that keeping her pillow and headboard apart is a means to separate her parents.[14] Accordingly, it is worth considering in some detail just what intentionalist account of symptoms and other wish-fulfilments might be suggested by Freud’s phraseology.

Magical Intentionalism

First, we can quickly reject as un-Freudian certain *possible* intentionalist accounts of wish-fulfilments because they do not do justice to the magical aspect on which Freud appears to lay emphasis. Thus, one could envision that the unconscious, prevented from seeking satisfaction for a libidinal wish unacceptable to consciousness, simply chooses to engage in some activity as a pleasurable, second-best substitute for the activity which would fully satisfy its wish, just as someone unable, say, to realize their dream of playing professional sports might find some compensation in competing on an amateur level. Of course, such an account

[13] Moreover, perhaps the separation itself could, in turn, be seen as a means to the further end of preventing her parents from having sex.

[14] However, I shall consider below whether Freud might not allow the possibility that the girl, in the very act of separating the pillow and headboard, actually believes that she is directly separating her parents, in view of the symbolic identifications she has made between those items and her father and mother. If so, then the form of magic that is most analogous to Freud’s view of symptoms, etc., may be *voodoo*.

confronts the question how symptoms, etc., which are at most symbolically or associatively connected with libidinal wishes are able to serve as pleasurable, second-best substitutes for the true objects of one's desire.[15] But leaving that question aside, such an account simply fails to do justice the fact that wish-fulfilments for Freud clearly lack the character of realistic satisfactions of desires, even of desires for things chosen as second-best substitutes for more ultimate objects of desire.[16]

More in keeping with Freud's phraseology is an account which sees wish-fulfilments as chosen as "magical" means to desired ends. Of course, as I noted, many recent commentators reject the intentionalistic reading of Freud which requires attributing to the unconscious bizarre beliefs that a given symptom constitutes the attainment of a (libidinal) wish or serves as a means to its attainment. Can Elektra, for example, really believe that by separating her pillow from the headboard she is preventing them from having sex? However, recall that Freud describes her as taking the pillow to have the meaning of a woman, the headboard that of a man. Perhaps what looks like a merely symbolic relation is actually something stronger: maybe the girl can be literally seen as unconsciously identifying the pillow with her mother and the headboard with her father. In that case, she would possess unconscious belief (or belief-like) states with the contents *pillow = mother* and *headboard = father*. In the presence of such identities, separating the pillow and headboard *constitutes* separating her parents, and so separating the pillow and headboard will appear a serviceable

[15] Can the unconscious, for example, be supposed to chose hysterical paralysis as a substitute for direct satisfaction of an Oedipal desire?

[16] Granted, I have suggested that Freud views the unconscious as sometimes reaching for realistic means to achieve desired ends, but the relevant phenomena should for that very reason not be accounted wish-fulfilments.

means to the end of preventing them from having sex. The girl's symptomatic act, then, can perhaps be viewed as performed on the model of the practical syllogism.[17,18]

Where would the odd identity-beliefs come from? Perhaps they could be seen as spontaneously generated by the id itself through association, insofar as they do not represent fixed, innate symbolic relationships. The ego would then be called upon to prevent the id from expressing contents unacceptable to conscious sensitivities. Censorship would be exercised by the ego as to which contents are fit to be expressed as symptoms, etc. Alternatively, suitable identity-beliefs could be seen as introduced by the censorship itself: Only those fit to generate symptoms, etc., sufficiently unobjectionable to conscious sensitivities need be introduced into the id. Granted,

[17] There are some complex questions about agency bound up with intentionalism. In performing her symptomatic act, the girl seems to have both a conscious intention that makes no reference to her parents and a conscious desire-belief pair that does. The question arises then which is her genuine intention in acting. Both 'intentions' would seem to have the appropriate logical relations to her action so as to rationalize it. Does the girl have two intentions in performing the act? Do the two intentions taken together constitute a compound intention of some sort? Perhaps the simplest picture of the situation is to see the conscious intention as the girl's genuine intention in performing the act: the contribution of the unconscious desire-belief pair would be limited to its part in producing the conscious intention. The story might go like this: The girl unconsciously wishes to separate her parents to keep them from having sex and believes that she can do so by separating pillow and headboard. So she unconsciously decides to implant the relevant desire into her consciousness. In that case, there *would* be an unconscious intention upstream of the girl's action but it would not be her intention in performing the action (in principle, however, even this element of unconscious intention might lapse from the account: the conscious intention to separate pillow and headboard could be viewed as a mere *causal consequence* of an unconscious desire to do so). Alternatively, though, it might be possible to see the *unconscious* intention as the girl's real intention in performing the action, with the conscious 'intention' as a mere epiphenomenon or confabulation. For example, the model of agency bound up in Rey's account of *akrasia* (see p. 117, n. 19 above) sees action quite generally as determined by one's unconscious beliefs and preferences (one's 'central' attitudes), with conscious beliefs and desires demoted to the status of mere 'avowed' attitudes. What holds of actions in general, would apply in particular to symptomatic acts like the girl's. I am not in a position to judge that any one of the variants of intentionalism just canvassed is more plausibly Freudian than the others. However, such details as those that divide the different versions may, as I discuss below, bear on whether and how wish-fulfilments turn out to be irrational.

[18] Perhaps it is not inapt to say that on the present account the id is susceptible to confusions between symbols and the things they symbolize. As a result, it can lead to an individual's behaving towards symbols as if they were their referents. It is tempting to suppose that voodoo involves a similar sort of confusion. Note, further, that Freud asserts that "Words were originally magic . . ." (1916-17, 17). Apparently, he identifies a primitive tendency to confuse symbols and things symbolized in a way that mirrors the operation of the id on the present account.

however one fills in the details, the resulting intentionalistic account may look extravagant. But as Freud himself suggests, neurotic symptoms are themselves often quite strange and may call for surprising explanatory hypotheses (1916-17, 268-69).

Intentionalism Applied to Arational Acts

It is interesting to consider whether something like the intentionalistic account of Freudian phenomena I have limned could apply, *mutatis mutandis*, to what are sometimes referred to as ‘arational’ (or ‘expressive’) actions, which bear some resemblance to symptomatic actions. Arational actions are supposed to be intentional actions that are not done for a reason, at least not one that involves the agent’s viewing the action as good in some respect(s), and which are typically explained through the agent’s being in the grip of some (strong) emotion (Hursthouse 1991). For example, an individual’s jumping for joy appears to be an intentional act that could well be performed without the individual’s believing that the jumping subserves some further aim or exhibits any particular good-making qualities at all. Everyday explanation of the act would confine itself typically to citing the agent’s strong feeling at the time: “I was so happy I just had to jump.”[19]

What is interesting in the present context about such arational actions[20] is that they often exhibit something like the symbolic or associative identifications that mark Freudian symptomatic acts (and other Freudian phenomena). For they often involve behavior directed towards items symbolic of, or associated with, the person (or thing) that is the object of one’s emotion. Thus, for example, someone enraged with an

[19] Or more fundamentally, an explanation could simply appeal to the agent’s desire to jump: “She just wanted to jump.” Reference to emotion, I think, could be seen as an indirect explanation of the action: it explains the desire that directly explains the action.

[20] The designation of such acts as ‘arational’ is Hursthouse’s. Her choice of this designation seems intended merely to highlight that such acts lack a classical rationalizing explanation in terms of a belief-desire pair. Ultimately, she avoids committing herself to the view that they lack reason-explanation altogether, let alone the view that they do not permit rational evaluation.

individual may destroy “her picture, letters or presents from her, awards from her, books or poems about her; the chair she was wont to sit in, locks of her hair, recordings [of] ‘our’ song, etc.” (Hursthouse 1991, 58). In such a case, it is tempting to suppose that one acts on such objects because it is impossible or imprudent to act on the object of one’s emotion itself. So perhaps there is some plausibility in seeing an individual’s destruction of a picture of his cheating wife, say, as a causal consequence of a desire to destroy the woman herself which is denied real satisfaction. Locating a desire behind the act in addition to the emotion felt begins to enlarge the analogy with Freudian symptomatic acts.

Of course, a big difference between such arational acts and symptomatic acts would be that in the former case there would be no reason to suppose that the relevant desire is unconscious, even temporarily. In many instances, for prudential reasons there may be a suppression of direct *action* towards the object of one’s emotion and desire, but there need be no suppression, let alone repression, of the emotion and desire themselves. So it is difficult to tell a story that explains arational acts precisely along the lines of the account of Freudian wish-fulfilments I have just described. The processes and acts the account refers to cannot, along with the pivotal desire, simply be transposed into consciousness with any plausibility. Certainly, the agent in performing an arational act has no awareness of actually forming an intention to, say, do violence to his cheating wife, let alone through tearing up her picture!

However, the fact that the pivotal desire is conscious does not necessarily preclude a role of the unconscious in generating arational acts. One way to accommodate the unconscious is along the lines of Rey’s distinction between avowed and central attitudes. On Rey’s account, one’s *conscious* desire to harm one’s cheating wife would amount to a mere avowed attitude, whereas the central attitudes genuinely responsible for one’s (non-verbal) actions would remain subconscious, even if some of them, in effect, amount to unconscious twins of one’s conscious attitudes. Thus, corresponding to one’s avowed desire to harm one’s cheating wife could very well be a central desire with the same content. Accordingly, the possibility opens up of that central desire’s combining with one’s unconscious, central beliefs in intentionalistic fashion to produce arational acts on the model of the practical syllogism. By analogy with the suggested intentionalistic account of Freudian phenomena above, one would simply need to suppose that a central belief arises temporarily whose content is *picture = wife*.

So it is possible to envision an account of at least some arational acts—namely, those with a seeming symbolic or associative character—that corresponds fairly closely to the intentionalistic account of Freudian phenomena.[21] The possibility of explaining arational acts in this way should reinforce one’s impression of the potential value of the intentionalistic account of Freudian phenomena itself as an explanatory hypothesis.[22] Greater clarity on the matter of intentionalism, however, will be possible after addressing another question important for the interpretation of

[21] I largely leave to one side the question how categories of arational acts that do not obviously partake of a symbolic character might be explained. However, I note that Hursthouse mentions a category of arational acts that are “explained by anger with inanimate objects” and consist in “doing things that might make sense if the things were animate,” for example, “kicking doors that refuse to shut and cars that refuse to start” (1991, 58). Once one allows the possibility of central beliefs as odd as those embodying symbolic identifications, one is unlikely to bridle at the possibility of (temporary) animistic central beliefs as well.

[22] The possibility of explaining some arational acts in this way belies Hursthouse’s assertion (1991, 63n6) that “the fascinatingly symbolic nature of many of the examples of arational action” does not “yield anything helpful” in reconciling such acts with the standard picture of intentional acts as permitting explanation in terms of belief-desire pairs. On the account just sketched, at least a subset of arational acts *do* permit such explanations, albeit in terms of an agent’s unconscious, central attitudes.

Freud’s view of symptoms, parapraxes, distorted dreams, etc.: Precisely in what sense does Freud take them to be *wish-fulfilments*?

The Nature of Wish-Fulfilments

In asking in what sense these phenomena are ‘wish-fulfilments’, the problem is not that which Freud himself raises with respect to distorted dreams, namely, that the existence of anxiety-dreams, in view of the conscious displeasure they occasion, seems to run counter to Freud’s claim that all dreams are wish-fulfilments. Freud solves *this* problem by consigning the wishes such dreams ‘fulfill’ and the pleasure they afford to the unconscious.[23] Rather, the problem that I want to highlight is more general, attaching to pleasant and unpleasant dreams alike, as well as neurotic symptoms, etc. Insofar as all these phenomena exhibit the distorting influence of dreamwork-processes, it is difficult to understand in what sense they can ‘fulfill’ wishes from whose contents they will typically starkly diverge. The difficulty is brought out by re-considering Wollheim’s account of the essence of wish-fulfilment (cf. p. 166, n. 12 above). Wollheim distinguishes between satisfaction and ‘gratification’ (ostensibly, ‘fulfilment’ in Freud’s sense) of a desire so:

> my desire that *p* is *satisfied* iff *p*
>
> my desire that *p* is *gratified* iff it is for me as if *p* (Wollheim 1979, 47)

[23] “No doubt,” Freud writes, “a wish-fulfilment must bring pleasure; but the question then arises ‘To whom?’.” Freud answers his own query: “a dreamer in his relation to his dream-wishes can only be compared to an amalgamation of two separate people . . .” (1916-17, 215-16).

But at first sight at least, it is difficult to see in what sense my dream that *q* involves (is?) a state of affairs in which 'it is for me as if *p*', where *q* and *p*—owing to the dreamwork—are very different contents.[24]

Now one interpretation of how symptoms, etc., can be wish-fulfilments, I think, can be quickly dismissed. As I noted, Freud stresses the plasticity of wishful impulses within primary process. He writes of the 'mobility of cathexes' whereby one libidinal wish within the id can surrender its intensity to another. This possibility, Freud maintains, explains how frustrated impulses need not always lead to illness: "if the satisfaction of one" of the sexual instincts "is frustrated by reality, the satisfaction of another can afford complete compensation" (1916-17, 345). Moreover, Freud thinks that within certain limits, sexual instincts "exhibit a large capacity for changing their object," a capacity which is notably evinced in the phenomenon of sublimation, where a (genetically related) aim is substituted that is no longer even explicitly sexual and, because socially approved, can be openly indulged. This may suggest that in non-basic wish-fulfilment the wish-fulfilling element consists in satisfaction of a *descendant* of an initial repressed libidinal wish.[25] So, for example, the woman's imperious desire to summon her maid to view the stained tablecloth might be seen as a causal descendant of an inadmissible wish that itself permits of straightforward satisfaction.

However, Freud emphasizes that the possibilities for such substitution are decidedly circumcised (uh . . . I mean circumscribed!), and that it is precisely in cases

[24] Similarly, with respect to neurotic symptoms, etc. In what sense, is it for the woman summoning her housemaid to the table with the stained tablecloth as if she were her former husband giving proof of his potency to their former chambermaid? Or in what sense is it for a patient with a hysterical paralysis as if he is satisfying some unconscious libidinal wish?

[25] Or, in the case of distorted dreams, in *hallucination* of the satisfaction of such descendant desires.

of pathology where such protections, particularly, in view of “libidinal fixations” (346), are least availing. So it would be a mistake to view neurotic symptoms, etc., on the model of sublimations as relatively straightforward satisfactions of substitute-desires.[26] Freud stresses that “symptoms offer nothing real in the way of satisfaction” (301) and, as noted above, Freud asserts that “an ordinary distorted dream,” for example, “is the disguised fulfilment of a repressed wish” (217). He does not maintain that it is the open fulfilment of a *derivative* of such a wish.[27]

Freud’s view, rather, *appears* to be that there is a layer of hallucinatory gratification underlying the less basic sorts of wish-fulfilment. This emerges from a telling passage in which Freud compares symptom-formation to dream-formation. With respect to the latter, he observes, “The dream proper, which has been completed in the unconscious and is the fulfilment of an unconscious wishful phantasy, is brought up against a portion of (pre)conscious activity which exercises the office of censorship and which, when it has been indemnified, permits the formation of the manifest dream as a compromise” (359-60) Thus, Freud evidently takes the notion of latent dream quite seriously, as an unconscious hallucinatory precursor to manifest dream. What is more, he posits the existence of a hallucinatory experience (like the latent dream) to which neurotic symptoms bear much the same relation as manifest

[26] There *may* be a place for desires that are derivative of the unconscious in symptomatic acts (like the woman’s above) which constitute an individual’s conscious motivations in performing the act (cf.p. 217, n. 17). But such acts will not be wish-fulfilments for Freud because they satisfy these superficial desires. And, in any case, such desires do not have any obvious role in symptoms (e.g., hysterical paralyses) which lack an act-like character.

[27] Because of their distinctive character, then, sublimations escape the rough classification of Freudian phenomena presented above into defense mechanisms and wish-fulfilments. Note, further, the contrast between Freud’s account of sublimation and the variety of intentionalism that regards symptoms, etc., as selected as pleasurable, second-best substitutes for the objects of one’s unconscious wishes (p. 215). Freud’s account of sublimations is not intentionalist since it makes no reference to unconscious choice or agency. Moreover, whereas in sublimation one’s initial unconscious wish is dissolved or transformed into a conscious successor, on the relevant intentionalism the initial unconscious wishes behind symptoms, etc., remain intact.

dream to latent dream: “the symptom emerges as a many-times-distorted derivative of the unconscious libidinal wish-fulfilment” (360). So the puzzle I raised as to in what sense symptoms, etc., constitute wish-fulfilments appears at least partly answered inasmuch as on Freud’s account their genesis lies in unconscious hallucinatory wish-fulfilment. This connects them with the sort of ‘gratification’, the ‘as if’-experience, that Wollheim identifies as central to wish-fulfilment.

But it leaves unanswered in what sense symptoms, etc., are *themselves* wish-fulfilments—rather than mere causal consequences of them—as Freud insists that they are.[28] One intriguing idea how symptoms, etc., might be assigned a more significant role within the less basic forms of wish-fulfilment comes from considering whether an account of pretend-play that Carruthers (2006) has developed, incorporating elements of Damasio’s theory of practical reasoning and motivation (Damasio 1994), can be adapted to fit key cases of wish-fulfilment. Though this approach, I think, ultimately fails to yield a satisfactory gloss on Freud’s view of symptoms, etc., as wish-fulfilments, it is worth considering, nonetheless, for the illumination it sheds on that view, and because the resultant theory, although not following Freud to the letter, represents a potentially valuable account of symptoms, etc., that retains much of the spirit of Freud’s account.

Wish-Fulfilments as Pretend-Play?

In the present context, what is interesting about Carruthers’ account of pretend-play is that it attributes something of a wish-fulfilling function to children’s pretence, and this raises the question whether the neurotic’s symptomatic acts (and

[28] For example, he writes, “the symptoms serve for the patients’ sexual satisfaction; they are a substitute for satisfaction of this kind, which the patients are without in their lives” (1916-17, 299). He certainly *seems* to mean by this more than just that the symptoms are side-effects of a process in which patients receive a substitute-satisfaction.

perhaps other characteristically Freudian phenomena) might be seen as wish-fulfilling along broadly similar lines. On Carruthers' account, a child's episode of, say, pretending to talk to Grandma on the telephone while using a banana as a prop is interpreted by her unconscious, belief-generating modules as an episode of talking to Grandma on the telephone. Her (unconscious) representation of herself as talking to her Grandma, in turn, is taken as input by her emotional systems, which leads to emotional rewards much like those which would ensue from the actual event. As Carruthers writes, "by representing what she is doing *as* talking to Grandma she is able to generate many of the same feelings and positive emotions as would be derived from the real thing. Although *she* isn't [consciously] fooled by her pretence into thinking that she is actually talking to Grandma, her emotional systems *are* so fooled, and respond accordingly" (2006, 297).[29]

In Carruthers' terminology, "the child's enactment of talking to Grandma *quasi-satisfies* her actual desire to talk to Grandma" (*ibid*.). Such quasi-satisfaction seems to correspond in all essentials to the sort of 'as if'-experience ('gratification') which Wollheim identifies as central to Freudian wish-fulfilment.[30] Accordingly, the

[29] On Carruthers' view, it is only the girl's emotional systems which are fooled into thinking that she is actually talking to Grandma. Other systems represent her as doing so only in a distinctive *suppositional* fashion.

[30] Carruthers is aware that his account requires some elaboration in order to serve as a general account of childrens' pretend-play. Thus, for example, he describes a case where a child pretends he is a dead cat, not because this promises to quasi-satisfy any desire he can plausibly be supposed to have, but because he anticipates that the display will amuse his audience (2006, 298). Again, a child may pleasurably pretend that his father is a monster, although such a state-of-affairs, if real, would occasion nothing but horror. Adapting an idea advanced by some aestheticians, Carruthers tentatively suggests that the child's pleasure in such a case may derive from his ability to control his emotional response: "For at any moment, by reminding himself that this is only a game, and that no one is really going to hurt him, he can close down the pretend inputs to his emotion systems, hence shutting down or modulating their response. And this might (in a different way) be pleasurable" (2006, 299). Here the presence of a background belief that the pretended state-of-affairs is not real is essential to the ability of the pretence to afford pleasure. However, the presence of a *conscious* belief that not *p* in no way

question arises whether a symptomatic act, say, though strictly a doing of *x*, might similarly give rise to an unconscious representation that one is doing *y*, which, fed to one's emotional systems, may lead to emotional rewards (unconscious pleasure) because one has a desire that *y*. Undoubtedly, there are significant differences between Carruthers' account of pretend-play and this account of less basic wish-fulfilment—including the fact that in the case of wish-fulfilment the relevant desire (or wish) is unconscious—but the latter would retain the general structure of the former.[31] Again, the merit of such an account would be that it assigns to the symptoms themselves an *active* role in the gratification of unconscious wishes, in a way that might warrant seeing them—as Freud evidently does—as *themselves* wish-fulfilments.[32]

However, an account of this sort might seem to confront problems as an account of non-basic Freudian wish-fulfilment. In the first place, it requires attributing a large degree of confusion to the unconscious in its interpretations of symptomatic acts. Even a case like that of the tablecloth-lady—which on its face looks favorable to the present account in view of the analogy between the symptomatic act itself and what, according to Freud, it represents—presents difficulties. For the account to work, the lady's unconscious must be seriously mistaken about such things as the time at which her action occurs, who she is

militates against the supposition that the child's emotional systems are privy to an *unconscious* belief-like representation that *p*.

[31] Presumably, in the case of wish-fulfilment, elaborations like those—mentioned in the previous note—required to deal with children's pretend-play in full generality will be unnecessary. The pretence involved will uniformly serve the purpose of quasi-satisfying unconscious desires.

[32] It is worth noting as well that Carruthers' account of pretend-play, more or less as initially formulated, might also serve as an explanation of 'arational acts' that partake of a symbolic character. Such an explanation might appear less extravagant than one framed on the model of magical intentionalism.

summoning to see the tablecloth, what sort of object the tablecloth is, and even who the agent of the action is, since on Freud's reading her symptomatic act represents her *husband's* own display of his potency! Furthermore, the feasibility of such an account appears even less in—typically Freudian—cases where the dreamwork avails itself of merely verbal, punning associative links between latent contents and their conscious representatives. So, for example, in his account of the Ratman (1909, 188-89), Freud describes how "[o]ne day while he was on summer holidays the idea suddenly occurred to him that he was too fat [German '*dick*'] and that he must *make himself slimmer*." Subsequently, he developed the practice of leaving the table before dessert and running up the mountainside in the hot summer sun. This symptomatic act is explained when it is revealed that the woman whom he loves had appeared at the same resort at which he was staying in the company of her cousin Richard, thereby evoking his jealousy. So Freud interprets the Ratman's compulsion to lose weight as a veiled fulfilment of the desire "to kill this Dick." That the dreamwork, according to Freud, makes use of such remote, "purely external associations" as the verbal one between Richard and fatness renders the degree of analogy between symptomatic act and what it represents exceedingly slight. It might seem farfetched indeed to view the Ratman's unconscious as so confused that it interprets his efforts to lose weight as attempts to kill said Richard!

A way beyond this impasse, however, may be suggested by considering Carruthers' account of pretend-play in a little more detail. Carruthers himself must confront the question how the mind-reading system comes to interpret a child's pretence that it is *x*-ing as an episode of *x*-ing when, in fact, it is not *x*-ing at all. The issue becomes particularly acute when one considers that a child's pretend-play includes episodes in which they pretend to be someone other than themselves (an astronaut, say), maybe even living at another time (like a cowboy).

Now in his account (2006, 299-300), Carruthers assigns the central role to a mind-reading system which, taking the non-conceptual perceptual input generated by the child's pretend-play as input, yields a conceptual interpretation of that act as output. Aware of the appearance of a mismatch between the input and the required output, Carruthers suggests that the mind-reading system will be assisted in bridging this gulf by visual and verbal imagery evoked in the course of the pretence.[33] Thus, as a child pretends to talk to her Grandma, she will generate imagery of her Grandma and telephones as well of the "words uttered or mentally rehearsed" during the episode (2006, 300). Perhaps, abetted by this information, the mind-reading system will be able to conjure up the required interpretation.

Carruthers mentions but rejects an account which assigns the mind-reading system "privileged access to the contents of the child's own intentions/action schemata that generate and guide her movements" (2006, 300). Carruthers is committed to a view of self-knowledge which, like Ryle's, largely assimilates first-person and third-person meta-cognition (see, e.g., Carruthers [2009b]; Ryle [1949]). On this view, one infers one's own mental states (and actions), just as one does another individual's, primarily on the basis of overt behavioral and contextual cues. Although Carruthers departs from Ryle in permitting the mind-reading system direct

[33] Carruthers note, further, "Like all other belief-generating systems, it [i.e., the mind-reading system] will be receiving its perceptual inputs 'tagged' for the pretend or suppositional status of the action. So this will be a crucial cue for the mind-reading system, telling it to set to work to interpret the perceived action as suppositional, searching for a content to attach to it that can make sense of it" (2006, 300).

access to content of an imagistic character, for various reasons he regards an architecture that credits mind-reading with access to intentions and other propositional attitudes as unlikely. However, in the present context—the adaption of Carruthers' model to symptoms, etc.—the sorts of empirical considerations that Carruthers would cite do not carry quite the same weight. Accordingly, an account of non-basic wish-fulfilment inspired by Carruther's account of pretend-play could posit a link, not between *conscious* intentions and mind-reader, but between *unconscious* intentions and mind-reading system. In that case, perhaps the typically quite large discrepancy between the content of the symptomatic act and its meaning, as it were, could be satisfactorily bridged.[34]

However, it's not clear how far this idea can be pushed as an interpretation of non-basic Freudian wish-fulfilment. As noted, part of the appeal of the present account is that it seems to assign symptoms, etc., a causal role in the gratification of unconscious wishes. But it seems implausible that the Ratman's mind-reading system, even supplied the content of his unconscious wish to kill Richard, can conjure up an 'as-if' experience corresponding to the content of that wish from the perceptual input generated by his weight-losing efforts. The mismatch between input and output appears too great. Indeed, even in cases where the discrepancy is considerably less—ones more like Carruthers' example of a child using a banana to pretend to talk to her

[34] Could the sorts of cues that Carruthers highlights—visual and verbal imagery, etc.— even be supposed to be available? Freudian theory's commitment to unconscious sensory imagery is fundamental; but it may a bit much to suggest that the thought processes of the unconscious are carried out to the regular accompaniment of verbal imagery. Granted, verbal associations, slips of the tongue, etc., loom large in Freud's view of the operation of the unconscious, but to admit that falls far short of crediting the unconscious with a kind of mental sub-vocalization. So, apart from the considerations mentioned in the text, it may be necessary to assign the mind-reader direct access to unconscious intentions in order to make up for the lack of such verbal cues.

Grandma—I am doubtful that the account can deliver anything phenomenally like an 'as if' experience.

Carruthers' view of perception lays great weight on the role of conceptualization. And it is true that, in a case like that of Wittgenstein's famous duck-rabbit (Wittgenstein 1958, 194), whether one perceives the figure as a duck or as a rabbit has much to do with the concepts that are activated while viewing it. Indeed, it even seems possible to determine one's perception by imposing one or the other concept on the figure in a top-down fashion. However, one would think that the part played by conceptualization in perception has its limits. Thus, though one can choose whether to perceive Wittgenstein's figure as either a duck or as a rabbit (by imposing the relevant concept), one cannot plausibly choose to perceive it as, say, a spider or starfish.[35] Presumably, this is because in such a case suitable non-conceptual sensory input is lacking. So, similarly, on the present account it is a stretch to hold that—at some unconscious level—it is for Elektra *as if* she were separating her parents in separating pillow from headboard .[36] Nonetheless, because a representation *is* delivered to her emotional systems with the content *separating mother and father*, her desire may count as being gratified, albeit in a non-phenomenal manner: her emotional systems can be supposed to respond with pleasure to the *information* that the desired state is realized. So although the present account

[35] To be sure, there *is* a sense in which one can see *x* as *y* where these items furnish utterly discrepant sensory input. Thus, for example, one can imagine an acting student given the exercise of seeing a coffee mug as a dagger. However, 'seeing as' in that sense just amounts to 'pretending that *x* is *y*'. On Carruthers' account of pretend-play (and similarly with respect to its Freudian application) the child *sees* the banana *as* a telephone in this sense, but her mind-reading system is supposed to *see* her *as* talking to Grandma in a more literal, quasi-perceptual sense. It is this latter sense that is at issue in the present discussion.

[36] Or even on Carruthers' original account of pretend-play that it is for the child *as if* she is talking to her Grandma.

of symptoms does fail to find room for one key element of Freudian wish-fulfilment, a robust 'as if' experience, it does correspond to Freudian theory in some other respects.

It may be a weakness of the present account, however—at least as a reconstruction of Freudian theory—that it would not obviously permit application to every class of phenomena which Freud regards as wish-fulfilling. Perhaps besides symptomatic acts the account could apply to distorted dreams, at least insofar as their manifest content represents the dreamer performing some action. The dream-imagery could conceivably be fed to the unconscious and thereby become available for (re-)interpretation by it. However, the current account would not seem to permit application at all to symptoms which lack an act-like character, for example, hysterical paralyses. For the account limits the role of unconscious interpretation to the assignment of content to consciously performed *actions*; such inert states as paralyses would seem to escape its scope.[37]

The Pretence Account vs. Magic Intentionalism

At this point, it is worth comparing and contrasting the model of wish-fulfilment suggested by Carruthers' account of pretend-play with the magic intentionalism that also seemed to correspond *in some respects* with Freud's pronouncements about wish-fulfilments. In the first place, the former account *would*

[37] Of course, on an intentionalist reading such manifestations as hysterical paralyses *are* actions (or at least the effects of actions). However, even on an intentionalist construction, they would seem to fall out of the scope of the present account, since outwardly they *appear* to lack the character of actions (indeed, for defensive purposes, it is crucial that they appear to lack that character). So they would not seem to provide the right sort of perceptual input to a mind-reading module.
Ultimately, however, I wish to leave open the possibility that the present account (or a modification of it) *might* be able to accommodate even manifestations that lack an explicit act-like character. Indeed, if the mind-reader performs its interpretative work more on the basis of accessed intentions than perceptual inputs, symptoms' lacking an explicit act-like character may pose no insuperable barrier to interpretation.

count as a form of intentionalism as well, since it would attribute the genesis of wish-fulfilments to unconscious intentions. But the content of the relevant intentions would differ from those of magic intentionalism. They would not be intentions to *actually* realize unconscious wishes but, rather, intentions to *pretend* that one is doing so: Elektra's episodes of separating pillow and headboard, for example, would be explained by an intention to simulate separating mother from father. Moreover, as a result, the account would seem to require complicating the structure of the unconscious. Whereas on magic intentionalism it is, ultimately, the same unconscious that both generates wish-fulfilling manifestations and interprets them as realizing one's unconscious wish, on the pretence account there would, it appears, need to be a split between the unconscious module or compartment generating the pretence and the unconscious compartment that receives the interpretation of the pretence. Otherwise it is hard to see how the unconscious would in any sense be 'fooled' by the pretence as required by the account.

Note, further, that in certain respects magic intentionalism more satisfactorily accords with Freud's characterization of non-basic wish-fulfilment. On magic intentionalism, the unconscious contents from which symptoms spring include odd identity beliefs, such as *pillow = mother* and *headboard = father*. So whereas the pretend-play account had some trouble explaining how the unconscious could be so confused as to interpret symptomatic acts like Elektra's as satisfying unconscious wishes, magic intentionalism provides a straightforward explanation. Since Elektra is supposed to have these odd identity beliefs, she can actually be held to experience her symptomatic act *as* a separating of her mother and father. She can be supposed to have the sort of 'as if' experience that appears to be central to Freud's understanding of wish-fulfilment.[38] Magic intentionalism, then, assigns a role to symptoms, etc., on which they can clearly count as wish-fulfilments themselves: they will be judged by the unconscious as satisfying its wishes and, therefore, can plausibly contribute to the wishes' quiescence.

However, though magic intentionalism yields a fairly satisfactory sense in which—at least some—symptoms, etc., might be seen as wish-fulfilments, it fails to do justice to the fact that Freud actually seems to locate the crucial element of gratification (the 'as if'-experience) *before* the symptoms, etc., in the causal chain, not after it as does the present account. As noted (p. 223 above), Freud appears to see unconscious hallucinatory wish-fulfilment as antecedent to both distorted dreams and symptoms, etc. In fact, the less basic forms of wish-fulfilment apparently presuppose the infantile hallucinatory form, with symptoms, etc., merely constituting an extra link inserted in the chain leading from wish through 'as if'-experience to the wish's quiescence. This circumstance, in fact, may push in the direction of a *non*-intentionalist reading of wish-fulfilments.

The Anti-Intentionalist Reading of Wish-Fulfilment

David Livingstone Smith explicitly recognizes the dependence of less basic forms of wish-fulfilment like symptomatic acts and parapraxes on the basic, hallucinatory form (Smith, 174-6). Discussing a parapraxis in which Freud forgets to send his proofs of his pamphlet "On Dreams"—according to Freud's self-interpretation, as a result of a wish that the work not be published—Smith offers the

[38] Even the Ratman can perhaps be supposed to experience his efforts to lose weight as killing Dick (see p. 227 above). For the identificatory beliefs magic intentionalism refers to could concern actions just as well as other categories. Thus, the Ratman could be supposed to have the belief *trimming = killing*.

following account: "(1) Freud wishes not to have 'On dreams' published (W_1), (2) W_1 conflicts with his [conscious] wish to have 'On dreams' published (W_2), (3) Freud repudiates W_1, (4) Freud unconsciously represents W_1 as fulfilled and (5) this brings about Freud's forgetting the proofs" (1999, 175-76). Essentially, on Smith's account, then, a frustrated wish leads to hallucinatory wish-fulfilment and, in turn, some sort of manifestation like a parapraxis or symptom.

But such an account does leave it somewhat puzzling in what sense symptoms, etc., are themselves wish-fulfilments. There is no easy answer to this question. Perhaps it suffices if they can be assigned *some* role in bringing about the quiescence of the wishes which instigate them. But it is left unclear *how* a symptomatic act, say, can lead to the quiescence of the wish that gave rise to it.[39] The account of symptoms, etc., based in Carruther's account of pretend-play had the merit of attempting to explain how in terms of intervening mechanisms: symptomatic act gives rise to an 'as if'-experience which, in turn, leads to the wish's quiescence (where this latter mechanism, though unconscious, is essentially a familiar one—we all know that imaginative episodes can lead to the quiescence of desires).[40] And magic intentionalism makes a similar appeal to symptoms' role in generating 'as if'-

[39] Of course, there is much else that remains unexplained on Smith's sort of account of wish-fulfilments. With respect to basic wish-fulfilment itself, Freud notes that he gives no account of a central element of the dream-work, namely, "the transformation of thoughts into a hallucinatory experience" (1916-17, 213). Moreover, specifically with respect to the less basic forms of wish-fulfilment, as Smith points out, there is the problem of explaining how hallucinatory wish-fulfilments give rise to such manifestations as parapraxes and symptoms (1999, 176). Of course, Freud's appeal to dream-work processes like condensation and displacement goes some way towards filling this gap, but not quite the whole distance. For example, there is the difficulty of explaining what Freud calls "the choice of neurosis," that is, of explaining just *what* pathology (or other manifestation) will result in a given case. Again, there is a question why such outward manifestations are required at all: Why does the underlying hallucinatory wish-fulfilment not suffice to appease the relevant unconscious wishes?

[40] Of course, I have raised some doubt that the pretend-play account can *actually* accommodate an 'as if' experience. But, even so, the account grants symptoms the sort of causal role in producing unconscious emotional rewards that justifies viewing them as wish-fulfilments.

experiences. But why on an account like Smith's should the Ratman's compulsive exercise, say, (temporarily) still his wish to kill his perceived rival Richard? Freud, however, fairly clearly takes the symptoms, etc., to have this role. And perhaps their possession of this role is ultimately enough to warrant Freud's view of such manifestations as wish-fulfilments.

Intentionalism or Anti-Intentionalism?

So where are we left in settling the issue between intentionalism and anti-intentionalism in the interpretation of Freud? With respect to wish-fulfilling phenomena, the weight of evidence seems to be somewhat on the side of anti-intentionalism. Magic intentionalism is supported by some of Freud's phraseology[41] and the fact that it provides a clear sense in which symptoms, etc., would be wish-fulfilling. However, that Freud apparently locates the crucial 'as if'-experience before symptoms, etc., in the causal chain provides strong support for a non-intentionalist reading, as does his insistence on the close analogy between symptom- and dream-formation: Freud seems to regard both symptom and manifest dream as generated from latent content by non-intentionalist dream-work *mechanisms*.[42] Moreover, as with the pretence-account (see p. 231 and n. 37), there may be some strain even in applying magic intentionalism to manifestations besides symptomatic acts. Recall Grünbaum's query: "Can the paranoiac be warrantedly said to have unconsciously *intended* his delusional persecutory thoughts and comportment to accomplish his erotic objectives?" (1984, 76-77). More particularly, what identity beliefs can the paranoiac be held to have such that he actually sees the very fact of his

[41] The absence of any such phraseology supporting the pretend-play variety of intentionalism, by contrast, casts doubt on it as a strict interpretation of Freudian theory. However, the account may retain interest as a charitable reconstruction of (parts of) Freudian theory.

[42] But see (Hart 1982) for an interpretation of even dreams as intentionalist for Freud.

having persecutory thoughts as realizing the states aimed at by the homosexual desires supposed to underlie his illness?[43] Accordingly, even the potential sphere of application of magic intentionalism may be confined to symptomatic acts.[44]

However, as I noted above (p. 210ff.), there are clear intentionalist elements in Freud's theory. Moreover, it is likely that Freud conceives of the operation of repression and resistance on the intentionalist model. Certainly, the sort of flexible behaviors that Freud thinks patients exhibit so as to sabotage their treatment and keep unsavory contents unconscious suggest the kind of means-end reasoning associated with intentional action.[45] Indeed, Freud sometimes views the very generation of a patient's illness *as a whole* on the intentionalist pattern, a fact which brings the wish-fulfilling symptoms bound up in the illness partly into the ambit of intentionalism (see p. 212 above). It is *possible*, in fact, that Freud quite generally sees non-basic wish-fulfilling phenomena as non-intentionalist mechanisms which are activated in the service of intentions of one sort or another. Though, ultimately, there is no very strong textual support for attributing this view to Freud, it cannot be altogether ruled out that he sees wish-fulfilments as activated precisely with the intention of affording pleasure, or at least reducing in some measure the displeasure occasioned by the

[43] Perhaps one could hold that where—as in this case—the symptoms are contentful states, the unconscious could see its desires as satisfied because, in virtue of possessing suitable identity beliefs, it is able to decode the content of those states by substituting identicals for identicals. But any such account would diverge from magic intentionalism, which sees the symptom itself (not its content) as furnishing the unconscious a seeming satisfaction of its desires. Moreover, this model apparently can be applied only to belief-like states such as delusions, not phobias, say, let alone to states such as hysterical paralyses which altogether lack content.

[44] Perhaps it would be possible to read Freud as intentionalist with respect to symptomatic acts, non-intentionalist with respect to other wish-fulfilling phenomena. Note that the textual evidence that supports magic intentionalism comes from passages where Freud treats of symptomatic acts specifically.

[45] Perhaps too the fact that Freud charges the (unconscious) *ego* with the execution of the defensive processes of repression and resistance supports an intentionalist reading of them. For at least with respect to its *conscious* aspect, the ego is understood as agentive in character.

frustration of unconscious wishes.[46] This variant of intentionalism, which I shall refer to as *hedonistic intentionalism*, has the merit of doing justice in some measure to the fact that Freud's theory seems to incorporate both intentionalist and non-intentionalist elements. Accordingly, I accord it considerable attention below.[47]

Ultimately, however, I shall take no definitive stand as to the precise extent of intentionalism in Freud's theory. Rather, I shall take into account both intentionalist and non-intentionalist readings in considering the degree of irrationality bound up in Freudian phenomena and its possible implications for normativist constraints.

Implications for Normativism

Immediate Difficulties for Normativism

Overall, there exists little consensus about what rational norms there are. And any clarity that exists about how to assess rationality when the mind is seen as a unified field of mental states evaporates when, as on the Freudian model, it is viewed as fundamentally divided. Should, for example, each mental compartment be viewed as separate for the purposes of rational assessment? Or should such assessment be made primarily with respect to the person as a whole? On the former view, the mere presence of inconsistent beliefs in a person, for example, will not count as irrational provided that those beliefs are spread around internally consistent mental compartments. Davidson's view, however, is clearly that inconsistency across compartments negatively impacts an individual's rationality (see p. 120 above), and,

[46] The idea would be that the whole process leading from unconscious wish to hallucinatory gratification to outward manifestation would be intentionally initiated. This would contrast with basic wish-fulfilment which I have presented as an entirely automatic process.

[47] The pretend-play account too can fairly readily put into a hedonistic mold, if the pretence is seen as undertaken for the sake of affording pleasure or reducing displeasure (though in Carruthers' original treatment of pretend-play the motivational story is more complicated). Consequently, most of the discussion of hedonistic intentionalism below will be seen to apply to the pretend-play account as well.

generally, he appears to take the view that rationality- (and charity-) standards apply in the first instance to persons, not their mental parts. Inasmuch as Davidson is the chief target of my anti-normativist argument, perhaps it will acceptable for me to take for granted his view of the matter and show that Freudian phenomena challenge charity principles on his own conception of rationality.

However, certain immediate difficulties for normativism fall out of that conception. Davidson himself explicitly embraces a *Principle of Total Evidence*, "which counsels an agent to accept the hypothesis supported by the totality of evidence he or she has" (2004b, 190). Moreover, presumably Davidson would accept an extension of this principle to the sphere of practical-reasoning along such lines as the following: an agent should choose that action which is supported by the totality of his or her desires (and relevant background beliefs).[48] However, in the nature of the case, compartmentalization such as one finds in Freud renders (abundant) beliefs and desires unavailable to conscious processes of theoretical and practical reasoning.[49] So a Freudian architecture virtually ensures abundant violation of Davidson's *Principle*;[50] and if one interprets that principle strongly so as to require that all potentially relevant beliefs or desires be *actually canvassed* in processes of reasoning, then the Freudian architecture *entails* the regular violation of the *Principle* in all processes of conscious reasoning.[51] [52] On either reading, then, Freudian theory

[48] I shall interpret the *Principle of Total Evidence* broadly so as to encompass both theoretical and practical reasoning.

[49] Conversely, conscious beliefs and desires may be rendered unavailable to *un*conscious processes of reasoning.

[50] In fact, the problem discussed here arises for virtually any modular model of central processes.

[51] Without the stipulation that potentially relevant attitudes be *actually* canvassed, perhaps a cognitive system could mostly observe the principle if it turned out that unconscious beliefs and desires, though relevant to inference and choice, rarely tipped the scales against inferences and choices supported by purely conscious attitudes.

appears to challenge Davidson's Competence Principle. For it difficult to see a system that is constructed so as to permit abundant (or regular) violations of a rational norm as embodying that norm as part of its competence. Moreover, the consequent irrationality cannot be easily seen as fitting the pattern required by Davidson's Compartment Principle. For the irrationality will not be owing to the active interference of one mental compartment with another, but merely to the inability of, say, the conscious compartment to access the states contained within unconscious compartment(s). To view this state of affairs as interference with the conscious compartment on the part of the unconscious one(s) would strain both the spirit and the letter of Davidson's Compartment Principle.

Consider, further, what additionally seems to follow on the strong reading of the *Principle of Total Evidence*. Given Freud's model, all conscious (and even unconscious) non-deductive inference and possibly *all* practical-reasoning will be rendered procedurally irrational through the failure to consult the totality of potentially relevant beliefs and desires in the cognitive system. Though Davidson's Threshold Principle is none too sharply formulated, the level of procedural irrationality dictated by a compartmental model like Freud's must cast substantial doubt on any Threshold Principle that, like Davidson's, requires that minds meet a high standard of overall rationality. This impression is only strengthened when one factors in the substantial *statal* irrationality likely to result from regular violation of the *Principle of Total Evidence*. Presumably, many beliefs will be formed and many actions taken which are irrational when judged in the light of the *entirety* of relevant

[52] I shall suggest below Freud holds that unconscious processes of reasoning are similarly blind to potentially relevant *conscious* beliefs and desires.

beliefs and desires possessed by the cognitive system.[53] In such circumstances, it becomes dubious to insist that the rationality in a cognitive system must greatly outweigh the irrationality. In short, the very *structure* of the mind on the Freudian model, then, already renders Davidsonian normativist principles suspect.

The Problems Considered in Greater Detail

But it is illuminating to consider a bit more concretely how Freudian phenomena aside from basic wish-fulfilments put pressure on normativist principles. First, note that the issue between intentionalism and non-intentionalism proves somewhat less significant in assessing the matter than one might have initially supposed. A non-intentionalist reading of central Freudian phenomena of wish-fulfilment *does* diminish the domain in which phenomena can be assessed with respect to their practical rationality or irrationality: Symptoms and other manifestations will be seen as generated by processes to which practical-rational norms simply do not apply. However, where the products of these processes are beliefs or emotions (as in delusions and phobias), they can still be assessed for their rational coherence with one's other attitudes; and the mechanisms which produce the beliefs can be assessed by epistemic standards of reliability. Moreover, in any case—as I suggested above—it is simply not plausible to deny an important place to intentionalism in Freud's theory: Even if wish-fulfilments can perhaps be interpreted non-intentionalistically, other elements of his theory, for example, repression and defense, should not. So there will be *at least some place* for assessments of practical rationality in considering the operations of the Freudian unconscious.

[53] A particularly crass illustration consists in cases where, say, conscious beliefs are permitted to form that directly contradict unconscious beliefs.

Non-Intentionalism Again

But consider the non-intentionalist reading of wish-fulfilments for a moment. Because on this account, the less basic forms of wish-fulfilment will include an undercurrent of basic, hallucinatory wish-fulfilment, the irrationality that attaches to the latter will attach to the former as well.[54] Accordingly, it appears that my argument against the Competence and Compartment Principles in the preceding chapter can also be made with the less basic forms of wish-fulfilment, if only because they include the basic form: they embody the same irrational competence as the basic form, in which the element of irrationality is owing to the id's internal operation, not interference by some other compartment of the mind.[55] Of course, there will be differences in the operation of wish-fulfilment subsequent to infancy. Because of the development of secondary process and the attendant expansion of opportunities for realistic satisfaction of one's desires, wish-fulfilment will be engaged with respect to a narrower set of desires, chiefly those consigned to the unconscious through repression. But when such desires are present, substitute-satisfaction *will* be sought for them. And Freud's view is that such desires are universally present, even in those not suffering from mental illness (hence his phrase "the psychopathology of everyday life"). In particular, during sleep, according to Freud, repressed desires find a wish-fulfilling outlet in the (distorted) dreams of normal and abnormal individuals alike. So wish-fulfilment remains a regular process even beyond infancy. Consequently, in

[54] I take for granted in the present discussion that basic wish-fulfilment is *ir*rational rather than *a*rational. See esp. p. 183ff above for defense of this assumption.

[55] The same points should hold on a hedonistic-intentionalist view of non-basic wish-fulfilments as well. For even if the sequence leading from unconscious wish to hallucinatory gratification (and, ultimately, to outward manifestation) is intentionally initiated (see p. 237, n.. 46 above), the hallucinations can be supposed to regularly give rise to corresponding irrational beliefs that one's wishes have been satisfied; and even the causal link between unconscious wish and hallucination can be seen as regular despite the intention that mediates them.

view of its undercurrent of hallucinatory wish-fulfilment and the resultant irrationality, non-basic wish-fulfilment challenges charity principles just as does the infantile form.

Less clear, however, is whether the added element of irrationality introduced by non-basic wish-fulfilments bears at all significantly on the Competence and Compartment Principles. Indeed, the irrationality bound up in neurotic symptoms, etc., on the Freudian conception seems, on its face, to fit the Davidsonian model expressed in the Compartment Principle fairly closely: irrational thoughts and behavior in one compartment of the mind (the ego) are produced by non-reason causal interference from another compartment (the id). Moreover, unlike the irrationality in basic wish-fulfilment, the outwardly irrational manifestations embodied in neurotic symptoms, etc., do not *clearly* reflect irrational competencies. Though there seems no principled reason that psychopathologies cannot be matters of competence rather than performance-error,[56] Freud is not sufficiently specific with respect to the aetiology of particular psychopathologies (and less basic wish-fulfilments generally) to allow one definitely to see them as reflecting *ceteris paribus* regularities of some sort.[57] Again, the very fact that they represent divergences from a norm is *prima facie* reason not to seem them as reflecting competencies pending countervailing evidence. In any case, Davidson's Competence Principle confronts no

[56] Perhaps the simplest way to make this possibility vivid is to envision a sub-population of human beings who are wired or programmed, as it were, in such a way as to regularly exhibit some pattern of deviant thought or behavior.

[57] To be clear: For Freud there *is* a *ceteris paribus* regularity that repressed desires will issue in some form of non-basic wish-fulfilment or other. But this falls short of an irrational competence. Indeed, the wish-fulfilling manifestations—as in the case of hysterical paralysis—need not even be rationality-evaluable. What one seems to need for an irrational competence, rather, are regularities linking particular kinds of repressed contents with particular kinds of irrational manifestations—and Freud does not plainly provide such regularities.

clear threat from the sorts of irrationality attaching distinctively to the less basic wish-fulfilments.[58]

Magic Intentionalism Again

Turning specifically to intentionalist accounts of Freudian phenomena, we see that additional potential sources of irrationality are created. Magic intentionalism, at the very least, introduces a significant element of *theoretical* irrationality into the Freudian unconscious. On the account, odd identity beliefs will be produced by (or introduced into) the id without being grounded in other belief(-like) states that could provide rational warrant for them; and, of course, the mechanisms responsible for their production can lay no claim to reliability.

Moreover, if one recognizes categories of conceptual truth and falsehood at all, surely many of the identities contemplated by the present account will be prime exemplars of the latter. Take, for example, the belief(-like) state that a pillow is one's mother. Conceptual incoherence may be lurking just below the surface.[59] However, even if one rejects conceptual truths and falsehoods, the introduction of such identities into the id would seem to render its contents logically inconsistent, at least to the extent that the id is supposed to partake of any significant portion of the realistic beliefs possessed by consciousness. So, for example, the belief that a pillow is one's mother introduced into a set of beliefs including the realistic beliefs that

[58] This general conclusion, however, may require some qualification with respect to dreams, which in Freud's view are generated as distorted wish-fulfilments from the latent content embodied in the underlying hallucinatory wish-fulfilment. For, to all appearances, in dreaming we temporarily come to accept altogether bizarre suppositions, even though those suppositions appear utterly senseless from the standpoint of other information possessed by our cognitive system (information which is likely to be accessed in other moments or episodes of dreaming). The regularity with which in dreaming such irrational suppositions are entertained perhaps *does* suggest an irrational competence.

[59] Or better still: Consider a man's belief that his genitals are identical to the number three (generated in virtue of an obvious association).

pillows are inanimate and that one's mother is animate forms a set of beliefs with logically inconsistent contents. In any case, an odd identity belief that a pillow is one's mother will straightforwardly clash with one's conscious belief that she is not. So the cognitive system as a whole will be mired in logical inconsistency. Again, if these odd identity-beliefs are seen as spontaneously generated by the id itself, then the theoretical irrationality associated with them will not be generated according to the model encapsulated in the Compartment Principle—reason once more to doubt the claim of that normativist principle to the status of necessary truth.

The magic-intentionalist model evinces *practical* irrationality inasmuch as the symptomatic acts which are produced according to it will, presumably, typically possess an akratic character. Accordingly, the resultant actions will generally violate the Principle of Total Evidence. Indeed, they will always violate the Principle in its strong form if in forming the relevant intentions involved the id routinely fails to consult possibly pertinent desires and information elsewhere in the cognitive system, in particular, within the conscious mind.[60] This failure itself, since not owing to extra-compartmental interference, would also represent a violation of the Compartment Principle.

Perhaps one will harbor doubt with respect to magic intentionalism whether it is really possible to harbor beliefs as odd as those the account contemplates. It is worth considering, though, how easily we entertain similarly bizarre hypotheses in the context of fiction and mythology: people being turned into frogs or even

[60] There would seem to be a failure, on the one hand, to weigh conscious desires and interests in selecting a course of action, and a neglect of consciously held beliefs in assessing whether such actions as moving pillows, etc., are serviceable means to the desired ends—a breach of instrumental rationality.

inanimate objects like mountains, rivers or what not; and, of course, these mythologies represent(ed) actual beliefs in their cultures of origin. Moreover, all manner of equally bizarre beliefs seem appropriately attributable to those suffering from psychotic delusions. Why should it not be possible, then, that some part of the neurotic (or maybe even the normal individual) is prone to similarly extravagant beliefs? Indeed, as noted above (p. 243, n. 58) magic intentionalism does not ascribe to the id much more in the way of surprising properties than seems to be involved in perfectly run-of-the-mill dreaming. So the account appears a coherent hypothesis that should not be ruled out of court.[61] Accordingly, the violations of charity principles bound up in it represent genuine problems for normativism.

[61] Doubt with respect to magic intentionalism may arise from another quarter, however: (1) If one unconsciously identifies pillow and mother, why does one treat the pillow as one's mother only in a very narrow range of behaviors? Why does one not, for example, invite it to Sunday brunch? And, (2), if by chance the pillow begins to lose its stuffing, is it for the id as if the same fate has befallen the mother? By way of response to (1), however, it suffices to observe that the id may harbor only very few mother-involving desires. Elektra's id, for example, may contain only (or chiefly) the desire to separate her mother from her father, resulting in her separating pillow and headboard. Other desires involving her mother would be conscious ones and, therefore, not suitably positioned to combine with the id's odd identity-beliefs in practical reasoning. With respect to (2), it is difficult to see how magic intentionalism can easily escape the implication that the id sees *whatever* happens to the pillow as equally happening to the mother. If so, this would seem to make the id emotionally hostage to the fate of the pillow, which might appear to detract somewhat from the id's function of affording unconscious pleasure. Perhaps, though, one can suppose the id's attention is confined solely to whatever is relevant to achieving its immediate narrow purposes.

At least this much force, however, must be conceded to the *first* objection: Magic intentionalism requires that the id be viewed as a separate center of agency (or influence on agency) from the mainstream of an individual's agency and practical reasoning. Thus, if one takes the view that action primarily emanates from one's unconscious central attitudes—as opposed to one's conscious avowed attitudes (see p. 117, n. 19 above)—one must see the id as a *distinct* influence on the agency associated with one's central attitudes, not as incorporated within it. In the same way, it seems one would need to distinguish the source of realistic Freudian intentionalistic behaviors from the source of magic-intentionalistic ones (*viz.*, the id). So some complexity in one's conception of the structure of the unconscious would be required—beyond that already dictated by the need to distinguish unconscious ego (the repressing force) from the id (the locus of repressed contents).

For a reading of Freud that, in essence, ascribes to him the distinction between central and avowed attitudes see Smith (1999, 162-66, esp. n163). Specifically, Smith introduces the distinction in explicating Freud's view of the *akrasia* bound up in symptomatic acts.

Note, further, the fact that at least some irrationality can in principle be explained via the discrepancy between avowed and central attitudes, rather than through compartmental interference, casts further doubt on Davidson's Compartment Principle (though, of course, on magic intentionalism the central attitudes themselves will be interfered with by the id).

Hedonistic Intentionalism

Whereas on magical intentionalism—with its actions based in bizarre identity beliefs—the irrationality of Freudian phenomena is fairly patent, on the hedonistic-intentionalist picture at first sight they offer some semblance of rationality. Is it not rational, say, for the id to seek some pleasurable gratification of wishes whose direct satisfaction is unavailable to it? Indeed, when interpreted on the hedonistic-intentionalistic picture, Freud can seem to present a picture of the mind as *hyper-rational*: not only do phenomena thought to be irrational (delusions, phobias, symptomatic acts, etc.) all receive coherent rational explanations but phenomena otherwise held to be altogether beyond the pale of rational assessment (hysterical symptoms, parapraxes, etc.) are brought within its scope. But the appearance of rationality is deceptive. To see this, consider once more how hedonistic intentionalism views Freudian phenomena.

Fundamentally, we have repression, resistance, and wish-fulfilling processes to consider. On hedonistic intentionalism, traumatic memories and unacceptable desires are consigned to the unconscious in order to prevent the pain which they would occasion the conscious mind. Once unconscious, resistance contrives expedients in order to neutralize forces—such as the therapist's probing—that threaten to raise those untoward contents to consciousness. Moreover, a wish-fulfilling substitute-satisfaction or gratification is sought for the unconscious wishes in order to diminish the pain of frustrated desire and afford some pleasure to the unconscious mind. In essence, a process is intentionally initiated that leads from unconscious wish to hallucinatory gratification, outward (or additional inward) manifestation, and, ultimately, the wish's relative quiescence. Is this unconscious pursuit of pleasure and pain-reduction not the very picture of rationality?

However, consider the theoretical (ir-)rationality bound up in this organization of the psyche. In the first place, the present reading does nothing to cancel the element of theoretical irrationality bound up in obsessive beliefs and other delusions, which are induced without proper regard to available evidence.[62] Moreover, there appears to be theoretical irrationality bound up in repression and resistance.[63]

Now repression of memories and the beliefs they involve does not quite seem to fit the usual understanding of self-deception, on which—roughly—"self-deception consists in getting yourself to believe one thing in order to avoid facing what you know to be the truth" (Gardner 1993, 16).[64] Repression, rather, consists in the *removal* of contents like beliefs from consciousness, not in the inducing of them. Yet it too is a motivated phenomenon and perhaps irrational in several respects. First, it creates what might appear to be a kind of rational incoherence, the simultaneous presence of a belief that *p* to the unconscious and its absence to consciousness. Perhaps a conscious-unconscious split is irrational by its very nature when it comes to

[62] The illusory beliefs arising in episodes of dreaming can count as further instances of theoretical irrationality, as can phobias if understood as on a cognitive theory of emotion.

[63] My remarks focus mainly on repression. However, several of the points apply, *mutatis mutandis*, to resistance as well.

[64] By contrast, resistance seems to operate in part precisely through self-deception. Thus, for example, Freud describes 'intellectual resistance', whereby patients induce in themselves doubts about the legitimacy of psychoanalytic theory and technique in order to escape the therapist's interpretations of their own case (1916-17, 289). However, Freud remarks, "The patient's resistance is of very many sorts, extremely subtle . . . and it exhibits protean changes in the forms in which it manifests itself " (287). For instance, Freud describes that in women resistance often exploits an erotic transference toward their doctor: "their jealousy . . . and their exasperation at their inevitable rejection . . . are bound to have a damaging effect on their personal understanding with the doctor . . ." (290-91). The unconscious is indeed a cunning little devil on the Freudian picture!

belief. [65] Second, at the very least, repression is likely to result in straightforward rational incoherence between conscious and unconscious, in that the conscious, deprived of relevant autobiographical memories, will often form beliefs that run counter to the unconsciously entertained beliefs.[66] So repression, even if not theoretically irrational itself, seems to (causally) contribute to irrational states-of-affairs. But, arguably, undertaking to do what can be expected to result in theoretically irrational states-of-affairs itself represents a form of theoretical irrationality.[67, 68] Moreover, it is repression that in large part produces the conditions responsible for the regular violation of the Principle of Total Evidence: by consigning beliefs to the unconscious, it ensures that the mind will fail to take into account some portion of potentially relevant beliefs in its conscious reasoning and will thereby fall afoul of Competence and Compartment Principles (see pp. 238-39 above)—further grounds perhaps for seeing repression itself as irrational.

Bear in mind, further, that Freud holds that such repression takes place on a massive scale. In particular, Freud holds that as the child enters ‘latency’ around its sixth year (a period of relative sexual dormancy), “The majority of experiences and mental impulses before the start of the latency period now fall victim to infantile amnesia—the forgetting . . . which veils our earliest youth to us and makes us strangers to it.” Freud remarks, “It is impossible to avoid a suspicion that the

[65] For the most part, I am agnostic about the irrationality of this and the other features of repression catalogued here. I merely wish to broach the *possibility* of their irrationality.

[66] Of course, the conscious will sometimes be abetted in forming these contrary beliefs by processes of self-deception, but those processes are distinct from the repression which precedes them.

[67] I am assuming here that the repressing ‘agency’ *can* expect the irrational consequences of its undertakings. But even if it cannot, it is plausible to attribute to the cognitive system as a whole beliefs which would entail the likelihood of those irrational consequences occurring.

[68] Similarly, perhaps resistance is theoretically irrational—among other reasons—because it consists in doing what can be expected to *sustain* theoretically irrational states-of-affairs.

beginnings of sexual life which are included in that period have provided the motive for its being forgotten—that this forgetting, in fact, is an outcome of repression" (1916-17, 326). Repression, then, operates on a large scale for Freud, in which case the amount of irrationality associated with it could turn out to be very considerable indeed.[69]

Granted the theoretical irrationality of Freudian phenomena on the hedonistic-intentionalist model, what of their practical rationality? Do they not at least well serve the practical interests of the individual? After all, the hedonistic-intentionalist model sees the behaviors resulting in breaches of theoretical rationality as undertaken precisely with the purpose of sparing the individual pain and affording them pleasure. However, the view of Freudian phenomena as practically rational does not withstand scrutiny. In the first place, by consigning beliefs and desires to the unconscious, repression renders them unavailable to conscious practical reasoning; that is, repression ensures that the practical analogue of the Principle of Total Evidence will be regularly violated.[70] Such regular violation of a practical-rational norm seems a violation of the Competence and Compartment Principles; and, as before, repression's rationality is itself perhaps called into question itself in view of its causal contribution to this state-of-affair.[71]

[69] I return to this point below in discussing the implications of hedonistic intentionalism for Davidson's Threshold Principle.

[70] Note that even though the repressed desires will, presumably, tend to be ones that it would be imprudent to act on, neglecting to consult them in conscious processes of practical reasoning would still constitute violation of a strong Principle of Total Evidence. Moreover, Freud himself takes the view that *some* repressed desires are of a kind that an individual would, in fact, be better off indulging (1910, 53-54). In any case, there is clear irrationality in failing to consult all *beliefs* potentially relevant to practical reasoning, of which there are sure to be some among repressed contents.

[71] As above, resistance is implicated too by *sustaining* the states-of-affairs that result in such irrationality. But perhaps question can be raised whether it is not sometimes practically rational to induce (or sustain) in oneself states that result in practical irrationality. However, even if it should turn

Indeed, on close inspection, the entire workings of the Freudian unconscious as pictured on the hedonistic-intentionalist model appear mired in irrationality. For although repression, resistance, and wish-fulfilment ostensibly aim at the avoidance of pain and the pursuit of pleasure, it is highly doubtful that they must succeed in maximizing the well-being of the individual. In fact, in cases of severe neuroticism—where these phenomena are perhaps most active—they seem to fail egregiously. Consider, for example, the case of the middle-aged woman suffering from delusions of jealousy described above (pp. 210-11). Freud writes of her that through her delusions she was "embittering her own life and the lives of her relatives," that "her happiness had been destroyed," and that her symptoms are "accompanied by intense subjective suffering and . . . threaten the communal life of a family" (1916-17, 248-50). In such circumstances, it is difficult to defend the rationality of the processes supporting her delusions.

Granted, it is foreseeable, not actual, outcomes on which attributions of practical (ir-)rationality depend, and prior to an illness, and perhaps even at its onset, beliefs cannot with any certainty be ascribed to an individual that entail the inadvisability of engaging in the repressions, wish-fulfilments, etc., destined to produce and manifest a severe neurosis.[72] Yet by the time a full-blown illness has wreaked havoc on one's life, the individual as a whole would seem to possess information that *does* entail the foolishness of persisting in these pathogenic

out that there are counterexamples to the principle that it is always irrational to do so, my subsequent discussion should make it clear that repression and resistance are not likely to constitute such exceptions.

[72] Even the individual versed in Freudian psychoanalytic theory is likely not to know the autobiographical particulars (the fixations, genetic predispositions, etc.) which, according to Freud, partly determine how repression and other Freudian processes will play out in their case! (But, then too, at least some of this information may be available to the unconscious.)

activities, let alone in undermining through resistance every effort to undo the conditions supporting one's neurosis. Indeed, the psychoanalytic *patient* will often exhibit something like an akratic relation to their pathology: They will consciously deplore the machinations of their unconscious, judge that it would be best if it (and they) did not engage in them, but find themselves powerless to oppose them.[73]

Of course, one may reasonably broach the possibility that the interests served by the Freudian phenomena, nonetheless, outweigh the ones against which those phenomena militate. Perhaps, despite appearances, the pleasure and desire-satisfaction afforded the patient through their illness (through wish-fulfilment, through escaping recognition of unsavory facts about oneself, etc.) is greater than that otherwise to be had. Ultimately, it is not incumbent upon me to insist that it does not. For my argument, the mere scientific possibility that it does not suffices. However, one does well to recall that wish-fulfilments afford no real desire-satisfaction, but merely serve the interests of reducing unconscious pain and promoting unconscious pleasure. Furthermore, the hypothesis being entertained becomes highly implausible when, as sometime happens, a patient's condition is severe enough to drive them to the brink of suicide.

All in all, on the hedonistic-intentionalistic model, the Freudian unconscious gives every appearance of being narrowly and inflexibly focused on the procurement of unconscious pleasure, with blithe neglect of the long-term, overall interests of the

[73] Of course, the condition described differs from ordinary *akrasia* in that the states constituting the condition are distributed among both conscious and unconscious compartments. It is not obvious, though, that this bears on the condition's irrationality.
Note that the state of the neurotic *not* convinced of the truth of psychoanalytic theory and its applicability to their case will not so closely resemble *akrasia*, since, though they will judge that they would be better off without their illness, they will lack a judgment to the effect that they would be better off not to induce it in themselves (better off not to repress the relevant pathogenic beliefs, resist the therapists efforts to bring them to consciousness, etc.).

individual. In a word, it appears practically irrational. Moreover, if—as seems plausible—these features are assumed to be endemic to unconscious agency of the Freudian sort, then the Freudian unconscious seems to embody an irrational competence (contra Davidson's Competence Principle); and since the irrationality involved is not owing to inter-compartmental interference, they represent a violation of Davidson's Compartment Principle as well.[74] It remains, however, to consider the bearing of the less basic manifestations of the Freudian unconscious on the Threshold Principle.

Threshold Principle

In the preceding chapter, it was argued that the Threshold Principle is challenged by the possibility that irrational, purely wish-fulfilling processes might predominate prior to the development of secondary process. However, the foregoing discussion suggests that Freudian theory, especially when interpreted on the hedonistic-intentionalistic model, implies that there is (or can be) a considerable amount of irrationality in the human being even subsequent to the development of the—ostensibly more rational—secondary process. As I have emphasized, the divorce between conscious and unconscious processes assures that most reasoning, both theoretical and practical, on the side of consciousness will be procedurally

[74] These conclusions could perhaps be resisted by urging that that unconscious agency functions quite differently in the case of the healthy and the severely ill. However, that suggestion seems to run counter to Freud's own conception of the matter. Generally, the unconscious would seem to pursue a narrow agenda while remaining blind or indifferent to the individual's long-term, overall interests. Though in the case of normal individuals this state-of-affairs may result in relatively favorable outcomes, even here the unconscious would exhibit procedural irrationality through failing to consult broader interests. That is, the unconscious appears to violate a strong reading of the Principle of Total Evidence.

irrational when assessed by standards of ideal rationality.[75, 76] Again, at least with respect to its practical reasoning, the Freudian unconscious, in turn, appears to fall afoul of those same norms in virtue of its narrow pursuit of pleasure via wish-fulfilment. So Freud's basic picture of the organization of the personality itself challenges the insistence on a high level of rationality as a necessary condition for agency.

Furthermore, the total amount of irrationality of the cognitive system increases when one factors in the additional irrational processes Freud postulates. Even in the relatively healthy individual, there will be the beliefs formed irrationally on the basis of hallucinatory wish-fulfilment and dreaming; and there will be parapraxes and transferences.[77] In the neurotic, added to this will be the assorted symptoms associated with their illness (symptomatic acts, delusions, phobias, etc.), as well as the irrational attempt to preserve the illness through resistance. Just how abundant these (irrational) neurotic manifestations can become according to Freudian theory is illustrated in case-studies like that of the Ratman (1909). So Freudian theory seems to contemplate the possibility of agents who very substantially fall short of the standard of rationality embodied in Davidson's Threshold Principle.

[75] Freud, of course, did not contemplate the possibility of a distinctively cognitive unconscious. However, the incorporation of it into the Freudian picture would not appear to relieve the overall irrationality of the psyche on the Freudian picture. For the processes of a purely cognitive unconscious would, no less than conscious processes, presumably lack access to information contained within the dynamic unconscious and so equally fall afoul of the Principle of Total Evidence (of course, if the cognitive unconscious is itself modular and encapsulated, this only compounds the extent of violation of the Principle).

[76] Moreover, as noted, incoherencies between conscious and unconscious autobiographical beliefs are sure to develop, particularly as the result of infantile amnesia.

[77] Transference consists in the irrational redirection of unconsciously held emotional attitudes towards important figures in one's childhood onto people encountered in one's adult life (notably, one's therapist, but others as well).
Note that parapraxes seem not to represent mere violations of *procedural* rationality in the practical sphere; at least *prima facie*, they often seem to fly in the face of the overall practical interests of the individual.

Indeed, perhaps Freudian theory even permits making a *degenerative* argument (cf. p. 172ff above) against Threshold Principles that set the standard of rationality lower than does Davidson's. For it appears possible to imagine an unfortunate individual who becomes progressively afflicted with more and more of the neuroses of which Freudian theory treats. In fact, for any threshold of requisite rationality which a normativist might care to set, we seem able to imagine this individual eventually so severely afflicted with these conditions and their attendant irrationality that they can be supposed to have fallen below that threshold—without thereby ceasing to be agents or bearers of propositional attitudes. So much the worse for the Threshold Principle.

All in all, the scientific possibility of the truth of Freudian hypotheses canvassed in the present chapter seems to provide ample grounds to call into question the *a priori* status of the charity principles championed by normativists of Davidson's stripe.

Concluding Thoughts

The upshot of the arguments of the preceding chapters is that the possibility of irrational subsystems of the mind puts normativism under severe strain. In Chapter Three, I presented the schema of an argument against normativism and proceeded in Chapters Four and Five to flesh out that schema with subsystems deriving from Freudian psychoanalytic theory. However, it is worth noting that subsystems posited in other reaches of psychological theory might well serve to put flesh on the bones of my argument.

For example, contemporary theorists posit an unconscious subsystem[78] of the mind (typically referred to as 'System1') dedicated to intuitive reasoning. This system is supposed to operate according to 'quick and dirty' heuristics that render it systematically prone to fallacies of reasoning (see, e.g., Carruthers [2009a], Kahneman [2002], Evans and Over [1996]). Since these divergences occur in virtue of the system's normal internal operations—and not through the external influence of some other mental compartment—it would seem to violate both the Competence and Compartment Principles. Moreover, inasmuch as many proponents of System1 reasoning view it as "a more primitive system present without the slower [conscious] one in many animals" (Rey 2007, 76), it appears to cast doubt on the Threshold Principle as well.[79]

However one fills out the argument-schema in detail, if done properly, the resulting argument should call into question both normativism and the hermeneutical conception of the mind insofar as it relies on normativism. If we wish to do justice to psychological phenomena, we should look elsewhere than to a philosophy of mind that misconstrues its fundamental character and places insufficiently motivated, artificial constraints on psychological theorizing.

[78] Or collection of subsystems.

[79] This is because in the absence of the—arguably—more rational processes of the conscious system ('System2') the irrational processes of System1 would predominate.

Bibliography

Antony, Louise. 1989. Anomalous monism and the problem of explanatory force. *Philosophical Review*, 98, no. 2 (April): 153-87.

Bermúdez, José. 2009. Nonconceptual mental content. *The Stanford encyclopedia of philosophy*. Fall 2009 ed. Ed. Edward N. Zalta. http://plato.stanford.edu/archives/fall2009/entries/content-nonconceptual/.

Boden, Margaret A. 1987. *Artificial intelligence and natural man*. 2nd ed, expanded. New York: Basic Books.

Bortolotti, Lisa. 2005. Delusions and the background of rationality. *Mind and Language*, 20, no. 2 (April): 189-208.

Brakel, Linda A. W. 2002. Phantasy and wish: A proper function account for human a-rational primary process mediated mentation. *Australasian Journal of Philosophy* 80, no. 1 (March): 1-16.

Carruthers, Peter. 2004. Suffering without subjectivity. *Philosophical Studies*, 121, no. 2 (November): 99-125.

------. 2006. *The architecture of the mind: Massive modularity and the flexibility of thought*. Oxford: Clarendon Press.

------. 2009a An architecture for dual reasoning. In *In two minds: Dual processes and beyond*, ed. J. Evans and K. Frankish. Oxford: Oxford Univ. Press.

------. 2009b. How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32, no. 2: 121-38.

Carruthers, Peter, and Peter K. Smith, eds. 1996. *Theories of theories of mind*. Cambridge: Cambridge Univ. Press.

Cavell, Marcia. 1986. Metaphor, dreamwork, and irrationality. In *Truth and interpretation: Perspectives on the philosophy of Donald Davidson*, ed. Ernest LePore, 495-507. Oxford: Basil Blackwell.

------. 1993. *The psychoanalytic mind: From Freud to philosophy*. Cambridge, MA: Harvard University Press.

------. 2002. Irrationality. In *The Freud Encyclopedia: Theory, therapy, and culture*, ed. Edward Erwin, 282-83. New York: Routledge.

Cherniak, Christopher. 1986. *Minimal* r*ationality*. Cambridge, MA: MIT Press.

Child, William. 1994. *Causality, interpretation, and the mind*. New York: Oxford Univ. Press.

Chomsky, Noam. 1980. *Rules and representations*. New York: Columbia Univ. Press.

Chomsky, Noam, and Jerrold Katz. 1974. What the linguist is talking about. *Journal of Philosophy* 71, no. 12 (June 27): 347-67. Quoted in Hookway 1988.

Cooper, W. E. 1980. Materialism and madness. *Philosophical Papers*, 9 (May): 36-40.

Cummins, Robert. 1983. *The nature of psychological explanation*. Cambridge, MA: MIT Press.

Damasio, A. R. 1994. *Descartes' error: emotion, reason, and the human brain*. New York: Avon Books.

Davidson, Donald. 1980a. Actions, reasons, and causes. In Davidson 1980b, 3-19.

------. 1980b. *Essays on action and events*. Oxford: Clarendon Press.

------. 1980c. Mental events. In Davidson 1980b, 207-27.

------. 1980d. Psychology as philosophy. In Davidson 1980b, 229-244.

------. 2001a. Belief and the basis of meaning. In Davidson 2001b, 141-54.

------. 2001b. *Inquiries into truth and interpretation*. 2nd ed. Oxford: Clarendon Press.

------. 2001c. Introduction. In Davidson 2001b, xv-xxiii.

------. 2001d. The method of truth in metaphysics. In Davidson 2001b, 199-214.

------. 2001e. Radical interpretation. In Davidson 2001b, 125-39.

------. 2001f. Theories of meaning and learnable languages. In Davidson 2001b, 3-15.

------. 2001g. Thought and talk. In Davidson 2001b, 155-70.

------. 2004a. Deception and division. In Davidson 2004d, 199-212.

------. 2004b. Incoherence and irrationality. In Davidson 2004d, 189-98.

------. 2004c. Paradoxes of irrationality. In Davidson 2004d, 169-87.

------. 2004d. *Problems of rationality*. Oxford: Clarendon Press.

------. 2004e. A unified theory of thought, meaning, and action. In Davidson 2004d, 151-66.

Dennett, D. C. 1969. *Content and consciousness*. New York: Humanities Press.

------. 1971. Intentional systems. *The Journal of Philosophy* 68, no. 4 (February 25): 87-106.

Erwin, Edward. 1996. *A final accounting: Philosophical and empirical issues in Freudian psychology*. Cambridge, MA: MIT Press.

Evans, Jonathan, and David E. Over. 1996. *Rationality and reasoning*. Hove, East Sussex: Psychology Press.

Evnine, Simon. 1991. *Donald Davidson*. Stanford: Stanford Univ. Press.

Flew, Anthony. 1956. Motives and the unconscious. In *Minnesota Studies in the Philosophy of Science, Vol. I*, ed. Herbert Feigl and Michael Scriven. Minneapolis, Univ. of Minnesota Press.

Fodor, J. A. 1974. Special sciences (Or: The disunity of science as a working hypothesis). *Synthese*, 28, no. 2 (October): 97-115.

------. 1983. *The modularity of mind*. Cambridge, MA: MIT Press.

Forbes, Graeme. 2009. Intensional Transitive Verbs. *The Stanford encyclopedia of philosophy*. Fall 2008 ed. Ed. Edward N. Zalta. http://plato.stanford.edu/archives/fall2008/entries/intensional-trans-verbs/.

Freud, Sigmund. 1953-74. *The standard edition of the complete psychological works of Sigmund Freud*. Trans. and ed. James Strachey. London: Hogarth Press.

------. 1909. Notes upon a case of obsessional neurosis. *S.E.* 10.

------. 1910. Five lectures on psychoanalysis. *S.E.* 11.

------. 1911. Formulations on the two principles of mental functioning. *S.E.* 12.

------. 1915. The unconscious. *S.E.* 14.

------. 1916-17 [1915-17]. *Introductory lectures on psycho-analysis*. *S.E.* 15 and 16.

------. 1923. *The ego and the id*. *S.E.* 19.

Gardner, Sebastian. 1993. *Irrationality and the philosophy of psychoanalysis*. Cambridge: Cambridge Univ. Press.

------. 1994. Psychoanalytic explanation. In *A companion to the philosophy of mind*, ed. Samuel Guttenplan, 493-500. Cambridge, MA: Blackwell.

Glock, Hans-Johann. 2003. *Quine and Davidson on language, thought and reality*. New York: Cambridge Univ. Press.

Goldman, Alvin. 1967. A causal theory of knowing. *Journal of Philosophy* 64, no. 12 (June 22): 357-72.

------. 1979. What is justified belief? In *Justification and knowledge*, ed. George Pappas, 1-23. Dordrecht: D. Reidel.

Gopnik, A. and A. Meltzoff. 1997. *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Grandy, Richard. 1973. Reference, meaning, and belief. *The Journal of Philosophy* 70, no. 14 (August 16), 439-52.

Grünbaum, Adolf. 1984. *The foundations of psychoanalysis: a philosophical critique*. Berkeley: Univ. of California Press.

Hall, Calvin S. 1954. *A primer of Freudian psychology*. New York: World Publishing.

Hart, W. D. 1982. Models of repression. In *Philosophical essays on Freud,* ed. Richard Wolheim and James Hopkins, 180-202. Cambridge: Cambridge Univ. Press.

Heil, John. 2004. *Philosophy of mind: A contemporary introduction*. 2nd ed. New York: Routledge.

Hempel, Carl G. 1966. *Philosophy of Natural Science*. Englewood Cliffs, NJ: Prentice-Hall.

Hookway, Christopher. 1988. *Quine: Language, experience and reality*. Stanford: Stanford Univ. Press.

Hornsby, Jennifer. 1997a. Introduction: Personal and subpersonal levels. In Hornsby 1997c, 157-67.

------. 1997b. Semantic innocence and psychological understanding. In Hornsby 1997c, 195-220.

------. 1997c. *Simple mindedness: In defense of naïve naturalism in the philosophy of mind*. Cambridge, MA: Harvard Univ. Press.

Hursthouse, Rosalind. 1991. Arational Actions. *The Journal of Philosophy* 88, no. 2 (February): 57-68.

Kahneman, Daniel. Maps of bounded rationality: A perspective on intuitive judgment and choice. Nobel Prize Lecture, 2002. Available at website: http://nobelprize.org/economics/laureates/2002/kahneman-lecture.html/.

Kihlstrom, J.F. 1999. The psychological unconscious. In *Handbook of personality*, ed. L.R. Pervin and O. John, 424-442. 2nd ed. New York: Guilford.

Kim, Jaegwon. 1993. *Supervenience and Mind: Selected Essays*. Cambridge: Cambridge Univ. Press.

Kripke, Saul A. 1980. *Naming and necessity*. Cambridge, MA: Harvard Univ. Press.

Lemmon, E. J. 1978. *Beginning logic*. Indianapolis: Hackett.

LePore, Ernest and Kirk Ludwig. 2005. *Donald Davidson : Meaning, truth, language, and reality*. Oxford: Oxford Univ. Press.

Lewis, David. 1972. Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy* 50: 249-58.

Ludwig, Kirk. 2004. Rationality, language, and the principle of charity. In *The Oxford handbook of rationality*, ed. Alred R. Mele and Piers Rawling, 343-60. New York: Oxford Univ. Press.

Marks, Charles E. 1981. *Commissurotomy, consciousness, and unity of mind*. Cambridge, MA: MIT Press.

McGinn, Colin. 1991. *The Problems of Consciousness*. Oxford: Blackwell. Quoted in Rey 2001.

McDowell, John. 1985. Functionalism and Anomalous Monism. In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, ed. Ernest LePore and Brian McLaughlin, 387-98. Oxford: Basil Blackwell.

Miller, Alexander. 1998. *Philosophy of Language*. Montreal: McGill's Univ. Press.

Millikan, Ruth. 1993. *White queen psychology and other essays for Alice*. Cambridge, MA: MIT Press.

Nagel, Ernest. 1979. *The Structure of Science: Problems in the Logic of Scientific Explanation*. Indianapolis: Hackett.

Phillips, James. 1996. Key concepts: Hermeneutics. *Philosophy, Psychiatry, & Psychology* 3, no. 1: 61-69.

Plato. 1961. *Phaedrus*. Trans. R. Hackforth. In *The collected dialogues of Plato including the letters*, ed. Edith Hamilton and Huntington Cairns, 475-525. Princeton, NJ: Princeton Univ. Press.

Pylyshyn, Zenon W. 2003. *Seeing and visualizing: It's not what you think.* Cambridge, MA. MIT Press.

Quine, Willard Van Orman. 1960. *Word and object*. Cambridge, MA: MIT Press.

------. 1976. Carnap and Logical Truth. In *The Ways of Paradox and Other Essays*, 2nd ed., 107-32. Cambridge, MA: Harvard Univ. Press.

------. 1980. Two Dogmas of Empiricism. In *From a logical point of view: Nine logico-philosophical essays*, 2nd ed., rev., 20-46. Cambridge, MA: Harvard Univ. Press.

------. 1986. *Philosophy of logic*. 2nd ed. Cambridge, MA: Harvard Univ. Press.

Rey, Georges. 1988. Toward a computational account of *akrasia* and self-deception. In *Essays in self-deception*, ed. Brian McLaughlin and Amélie Rorty, 264-95. Berkeley: Univ. of California Press.

------. 1994. Dennett's unrealistic psychology. *Philosophical Topics* 22, no. 1 & 2 (Spring and Fall): 259-89.

------. 1997. *Contemporary philosophy of mind*. Cambridge, MA: Blackwell Publishers.

------. 2001. Physicalism and psychology: A plea for a substantive philosophy of mind. In *Physicalism and its discontents*, ed. Carl Gillett and Barry Loewer, 99-128. New York: Cambridge Univ. Press.

------. 2007. Resisting normativism in psychology. In *Contemporary debates in philosophy of mind*, ed. Jonathan Cohen and Brian McLaughlin, 69-84. Oxford: Blackwell.

Ryle, Gilbert. 1949. *The Concept of the Mind*. New York: Barnes & Noble.

Samuels, R. 1998. Evolutionary psychology and the massive modularity hypothesis. *British Journal for the Philosophy of Science*, 49, no. 4 (December): 575–602.

Searle, John R. 1958. Proper names. *Mind* 67, no. 266 (April): 166-73.

------. 1992. *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Segal, Gabriel. 1996. The modularity of theory of mind. In *Theories of theories of mind*, ed. Peter Carruthers and Peter K. Smith, 141-157. Cambridge: Cambridge Univ. Press.

Smith, David Livingstone. 1999. *Freud's philosophy of the unconscious*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Soble, Alan. 1989. An introduction to the philosophy of love. In *Eros, agape, and philia: Readings in the philosophy of love*, ed. Alan Soble, xi-xxv.

Stein, Edward. 1996. *Without good reason: The rationality debate in philosophy and cognitive science*. New York: Oxford Univ. Press.

Stich, Stephen. 1978. Beliefs and subdoxastic states. *Philosophy of Science*, 45: 499-518.

------. 1983. *From folk psychology to cognitive science: The case against belief.* Cambridge, MA: MIT Press.

------. 1990. *The fragmentation of reason: Preface to a pragmatic theory of cognitive evaluation*. Cambridge, MA: MIT Press.

Suppes, Patrick, and Hermine Warren. 1980. On the generation and classification of defence mechanisms. In *Philosophical Essays on Freud*, ed. Richard Wollheim and James Hopkins, 163-79. Cambridge: Cambridge Univ. Press.

Toulmin, Stephen. 1954. The logical status of psycho-analysis. In *Philosophy and analysis,* ed. Margaret Macdonald. Oxford: Basil Blackwell.

Wegman, Cornelius. 1985. *Psychoanalysis and Cognitive Psychology*. London: Academic Press.

Wilson, N. L. 1959. Substance without substrata. *Review of Metaphysics* 12, no. 4 (June): 521-39.

Wilson, Timothy D. 2002. *Strangers to ourselves: Discovering the Adaptive Unconscious*. Cambridge, MA: Harvard Univ. Press.

Wittgenstein, Ludwig. 1958. *Philosophical investigations*. 3rd ed. Trans. G. E. M. Anscombe. Macmillan: New York.

------. 1961. *Tractatus logico-philosophicus*. Trans. D. F. Pears and B. F. McGuinness. London: Routledge & Kegan Paul.

Wollheim, Richard. 1979. Wish-fulfilment. In *Rational action: Studies in philosophy and social science,* ed. Ross Harrison, 47-60. Cambridge: Cambridge Univ. Press.